

# **NEW DEVELOPMENTS in COMPUTATIONAL INTELLIGENCE and COMPUTER SCIENCE**

**Proceedings of the International Conference on Applied Physics,  
Simulation and Computers (APSAC 2015)**

**Proceedings of the International Conference on Neural Networks -  
Fuzzy Systems (NN-FS 2015)**

**Vienna, Austria  
March 15-17, 2015**

# **NEW DEVELOPMENTS in COMPUTATIONAL INTELLIGENCE and COMPUTER SCIENCE**

**Proceedings of the International Conference on Applied Physics,  
Simulation and Computers (APSAC 2015)**

**Proceedings of the International Conference on Neural Networks -  
Fuzzy Systems (NN-FS 2015)**

**Vienna, Austria  
March 15-17, 2015**

**Copyright © 2015, by the editors**

All the copyright of the present book belongs to the editors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the editors.

All papers of the present volume were peer reviewed by no less than two independent reviewers. Acceptance was granted when both reviewers' recommendations were positive.

Series: Recent Advances in Computer Engineering Series | 28

ISSN: 1790-5109

ISBN: 978-1-61804-286-6

# **NEW DEVELOPMENTS in COMPUTATIONAL INTELLIGENCE and COMPUTER SCIENCE**

**Proceedings of the International Conference on Applied Physics,  
Simulation and Computers (APSAC 2015)**

**Proceedings of the International Conference on Neural Networks -  
Fuzzy Systems (NN-FS 2015)**

**Vienna, Austria  
March 15-17, 2015**



## Organizing Committee

### Editors:

Professor Yingxu Wang, Schulich School of Engineering, University of Calgary  
Professor Pierre Borne, Ecole Centrale de Lille, France  
Professor Imre Rudas, Obuda University, Budapest, Hungary

### Organizing Committee:

Prof. Panos M. Pardalos, University of Florida, USA  
Prof. Charalambos Arapatsakos, University of Thrace, Greece  
Prof. Maria Isabel Garcia-Planas, University Politecnica de Catalunya, Barcelona, Spain  
Prof. Anca Croitoru, Al.I. Cuza University, Iasi, Romania  
Prof. Aida Bulucea, University of Craiova, Craiova, Romania  
Prof. Nikos Mastorakis, Technical University of Sofia, Bulgaria  
Prof. Klimis Ntalianis, Technological Educational Institute of Athens, Greece

### Steering Committee:

Prof. Yuriy S. Shmaliy, IEEE Fellow, Universidad de Guanajuato, Mexico  
Prof. Alaa Khamis, IEEE Robotics and Automation Egypt-Chapter Chair, Egypt  
Prof. Eduardo Mario Dias, University of Sao Paulo, Brazil  
Prof. Miroslav Voznak, VSB-Technical University of Ostrava, Czech Republic  
Prof. Abdel-Badeeh M. Salem, Ain Shams University, Cairo, Egypt  
Prof. Antoanela Naaji, Vasile Goldis Western University Arad, Romania  
Prof. Elena Zamiatina, Perm State University, Perm Krai, Russia  
Prof. Pan Agathoklis, University of Victoria, Canada  
Prof. Claudio Talarico, Gonzaga University, Spokane, USA

### Program Committee:

Prof. Lotfi Zadeh (IEEE Fellow, University of Berkeley, USA)  
Prof. Leon Chua (IEEE Fellow, University of Berkeley, USA)  
Prof. Michio Sugeno (RIKEN Brain Science Institute (RIKEN BSI), Japan)  
Prof. Dimitri Bertsekas (IEEE Fellow, MIT, USA)  
Prof. Demetri Terzopoulos (IEEE Fellow, ACM Fellow, UCLA, USA)  
Prof. Georgios B. Giannakis (IEEE Fellow, University of Minnesota, USA)  
Prof. Abraham Bers (IEEE Fellow, MIT, USA)  
Prof. Brian Barsky (IEEE Fellow, University of Berkeley, USA)  
Prof. Aggelos Katsaggelos (IEEE Fellow, Northwestern University, USA)  
Prof. Josef Sifakis (Turing Award 2007, CNRS/Verimag, France)  
Prof. Hisashi Kobayashi (Princeton University, USA)  
Prof. Kinshuk (Fellow IEEE, Massey Univ. New Zeland),  
Prof. Leonid Kazovsky (Stanford University, USA)  
Prof. Narsingh Deo (IEEE Fellow, ACM Fellow, University of Central Florida, USA)  
Prof. Kamisetty Rao (Fellow IEEE, Univ. of Texas at Arlington, USA)  
Prof. Anastassios Venetsanopoulos (Fellow IEEE, University of Toronto, Canada)  
Prof. Steven Collicott (Purdue University, West Lafayette, IN, USA)  
Prof. Nikolaos Paragios (Ecole Centrale Paris, France)  
Prof. Nikolaos G. Bourbakis (IEEE Fellow, Wright State University, USA)  
Prof. Stamatios Kartalopoulos (IEEE Fellow, University of Oklahoma, USA)  
Prof. Irwin Sandberg (IEEE Fellow, University of Texas at Austin, USA),  
Prof. Michael Sebek (IEEE Fellow, Czech Technical University in Prague, Czech Republic)  
Prof. Hashem Akbari (University of California, Berkeley, USA)  
Prof. Lei Xu (IEEE Fellow, Chinese University of Hong Kong, Hong Kong)  
Prof. Paul E. Dimotakis (California Institute of Technology Pasadena, USA)  
Prof. Martin Pelikan (UMSL, USA)

Prof. Patrick Wang (MIT, USA)  
Prof. Wasfy B Mikhael (IEEE Fellow, University of Central Florida Orlando, USA)  
Prof. Sunil Das (IEEE Fellow, University of Ottawa, Canada)  
Prof. Nikolaos D. Katopodes (University of Michigan, USA)  
Prof. Bimal K. Bose (Life Fellow of IEEE, University of Tennessee, Knoxville, USA)  
Prof. Janusz Kacprzyk (IEEE Fellow, Polish Academy of Sciences, Poland)  
Prof. Sidney Burrus (IEEE Fellow, Rice University, USA)  
Prof. Biswa N. Datta (IEEE Fellow, Northern Illinois University, USA)  
Prof. Mihai Putinar (University of California at Santa Barbara, USA)  
Prof. Włodzisław Duch (Nicolaus Copernicus University, Poland)  
Prof. Michael N. Katehakis (Rutgers, The State University of New Jersey, USA)  
Prof. Pan Agathoklis (Univ. of Victoria, Canada)  
Dr. Subhas C. Misra (Harvard University, USA)  
Prof. Martin van den Toorn (Delft University of Technology, The Netherlands)  
Prof. Malcolm J. Crocker (Distinguished University Prof., Auburn University, USA)  
Prof. Urszula Ledzewicz, Southern Illinois University, USA.  
Prof. Dimitri Kazakos, Dean, (Texas Southern University, USA)  
Prof. Ronald Yager (Iona College, USA)  
Prof. Athanassios Manikas (Imperial College, London, UK)  
Prof. Keith L. Clark (Imperial College, London, UK)  
Prof. Argyris Varonides (Univ. of Scranton, USA)  
Dr. Michelle Luke (Univ. Berkeley, USA)  
Prof. Patrice Brault (Univ. Paris-sud, France)  
Prof. Jim Cunningham (Imperial College London, UK)  
Prof. Philippe Ben-Abdallah (Ecole Polytechnique de l'Université de Nantes, France)  
Prof. Ichiro Hagiwara, (Tokyo Institute of Technology, Japan)  
Prof. Akshai Aggarwal (University of Windsor, Canada)  
Prof. Ulrich Albrecht (Auburn University, USA)  
Prof. Alexey L Sadovski (IEEE Fellow, Texas A&M University, USA)  
Prof. Amedeo Andreotti (University of Naples, Italy)  
Prof. Ryszard S. Choras (University of Technology and Life Sciences Bydgoszcz, Poland)  
Prof. Remi Leandre (Université de Bourgogne, Dijon, France)  
Prof. Moustapha Diaby (University of Connecticut, USA)  
Prof. Brian McCartin (New York University, USA)  
Prof. Anastasios Lyrantzis (Purdue University, USA)  
Prof. Charles Long (Prof. Emeritus University of Wisconsin, USA)  
Prof. Marvin Goldstein (NASA Glenn Research Center, USA)  
Prof. Ron Goldman (Rice University, USA)  
Prof. Ioannis A. Kakadiaris (University of Houston, USA)  
Prof. Richard Tapia (Rice University, USA)  
Prof. Milivoje M. Kostic (Northern Illinois University, USA)  
Prof. Helmut Jaberg (University of Technology Graz, Austria)  
Prof. Ardeshir Anjomani (The University of Texas at Arlington, USA)  
Prof. Heinz Ulbrich (Technical University Munich, Germany)  
Prof. Reinhard Leithner (Technical University Braunschweig, Germany)  
Prof. M. Ehsani (Texas A&M University, USA)  
Prof. Sesh Commuri (University of Oklahoma, USA)  
Prof. Nicolas Galanis (Université de Sherbrooke, Canada)  
Prof. Rui J. P. de Figueiredo (University of California, USA)  
Prof. Hiroshi Sakaki (Meisei University, Tokyo, Japan)  
Prof. K. D. Klaes, (Head of the EPS Support Science Team in the MET Division at EUMETSAT, France)  
Prof. Emira Maljevic (Technical University of Belgrade, Serbia)  
Prof. Kazuhiko Tsuda (University of Tsukuba, Tokyo, Japan)  
Prof. Nobuoki Mano (Meisei University, Tokyo, Japan)  
Prof. Nobuo Nakajima (The University of Electro-Communications, Tokyo, Japan)

Prof. P. Vanderstraeten (Brussels Institute for Environmental Management, Belgium)  
Prof. Annaliese Bischoff (University of Massachusetts, Amherst, USA)  
Prof. Fumiaki Imado (Shinshu University, Japan)  
Prof. Sotirios G. Ziaavras (New Jersey Institute of Technology, USA)  
Prof. Marc A. Rosen (University of Ontario Institute of Technology, Canada)  
Prof. Thomas M. Gattton (National University, San Diego, USA)  
Prof. Leonardo Pagnotta (University of Calabria, Italy)  
Prof. Yan Wu (Georgia Southern University, USA)  
Prof. Daniel N. Riahi (University of Texas-Pan American, USA)  
Prof. Alexander Grebennikov (Autonomous University of Puebla, Mexico)  
Prof. Bennie F. L. Ward (Baylor University, TX, USA)  
Prof. Guennadi A. Kouzaev (Norwegian University of Science and Technology, Norway)  
Prof. Geoff Skinner (The University of Newcastle, Australia)  
Prof. Hamido Fujita (Iwate Prefectural University(IPU), Japan)  
Prof. Francesco Muzi (University of L'Aquila, Italy)  
Prof. Claudio Rossi (University of Siena, Italy)  
Prof. Sergey B. Leonov (Joint Institute for High Temperature Russian Academy of Science, Russia)  
Prof. Lili He (San Jose State University, USA)  
Prof. M. Nasseh Tabrizi (East Carolina University, USA)  
Prof. Alaa Eldin Fahmy (University Of Calgary, Canada)  
Prof. Gh. Pascovici (University of Koeln, Germany)  
Prof. Pier Paolo Delsanto (Politecnico of Torino, Italy)  
Prof. Radu Munteanu (Rector of the Technical University of Cluj-Napoca, Romania)  
Prof. Ioan Dumitrache (Politehnica University of Bucharest, Romania)  
Prof. Miquel Salgot (University of Barcelona, Spain)  
Prof. Amaury A. Caballero (Florida International University, USA)  
Prof. Maria I. Garcia-Planas (Universitat Politecnica de Catalunya, Spain)  
Prof. Petar Popivanov (Bulgarian Academy of Sciences, Bulgaria)  
Prof. Alexander Gegov (University of Portsmouth, UK)  
Prof. Lin Feng (Nanyang Technological University, Singapore)  
Prof. Colin Fyfe (University of the West of Scotland, UK)  
Prof. Zhaohui Luo (Univ of London, UK)  
Prof. Wolfgang Wenzel (Institute for Nanotechnology, Germany)  
Prof. Weilian Su (Naval Postgraduate School, USA)  
Prof. Phillip G. Bradford (The University of Alabama, USA)  
Prof. Hamid Abachi (Monash University, Australia)  
Prof. Josef Boercsoek (Universitat Kassel, Germany)  
Prof. Eyad H. Abed (University of Maryland, Maryland, USA)  
Prof. Andrzej Ordys (Kingston University, UK)  
Prof. T Bott (The University of Birmingham, UK)  
Prof. T.-W. Lee (Arizona State University, AZ, USA)  
Prof. Le Yi Wang (Wayne State University, Detroit, USA)  
Prof. Oleksander Markovskyy (National Technical University of Ukraine, Ukraine)  
Prof. Suresh P. Sethi (University of Texas at Dallas, USA)  
Prof. Hartmut Hillmer(University of Kassel, Germany)  
Prof. Bram Van Putten (Wageningen University, The Netherlands)  
Prof. Alexander Iomin (Technion - Israel Institute of Technology, Israel)  
Prof. Roberto San Jose (Technical University of Madrid, Spain)  
Prof. Minvydas Ragulskis (Kaunas University of Technology, Lithuania)  
Prof. Arun Kulkarni (The University of Texas at Tyler, USA)  
Prof. Joydeep Mitra (New Mexico State University, USA)  
Prof. Vincenzo Niola (University of Naples Federico II, Italy)  
Prof. S. Y. Chen, (Zhejiang University of Technology, China and University of Hamburg, Germany)  
Prof. Duc Nguyen (Old Dominion University, Norfolk, USA)  
Prof. Tuan Pham (James Cook University, Townsville, Australia)

Prof. Jiri Klima (Technical Faculty of CZU in Prague, Czech Republic)  
Prof. Rossella Cancelliere (University of Torino, Italy)  
Prof. Wladyslaw Mielczarski (Technical University of Lodz, Poland)  
Prof. Ibrahim Hassan (Concordia University, Montreal, Quebec, Canada)  
Prof. Erich Schmidt (Vienna University of Technology, Austria)  
Prof. James F. Frenzel (University of Idaho, USA)  
Prof. Vilem Srovnal, (Technical University of Ostrava, Czech Republic)  
Prof. J. M. Giron-Sierra (Universidad Complutense de Madrid, Spain)  
Prof. Rudolf Freund (Vienna University of Technology, Austria)  
Prof. Alessandro Genco (University of Palermo, Italy)  
Prof. Martin Lopez Morales (Technical University of Monterey, Mexico)  
Prof. Ralph W. Oberste-Vorth (Marshall University, USA)  
Prof. Photios Anninos, Democritus University of Thrace, Greece

### **Additional Reviewers**

Prof. Abelha Antonio, Universidade do Minho, Portugal  
Prof. Alejandro Fuentes-Penna, Universidad Autónoma del Estado de Hidalgo, Mexico  
Prof. Ana Maria Tavares Martins, University of Beira Interior, Portugal  
Prof. Andrey Dmitriev, Russian Academy of Sciences, Russia  
Prof. Angel F. Tenorio, Universidad Pablo de Olavide, Spain  
Prof. Athanassios Stavrakoudis, University of Ioannina, Greece  
Prof. Audenaert Amaryllis, Universiteit Antwerpen, Belgium  
Prof. Bazil Taha Ahmed, Universidad Autonoma de Madrid, Spain  
Prof. Bruno Marsigalia, University of Cassino and Southern Lazio, Italy  
Prof. Carla Falugi, University of Genova, Italy  
Prof. Carlos Gonzalez, University of Castilla-La Mancha, Spain  
Prof. Carlos Manuel Travieso-Gonzalez, University of Las Palmas de Gran Canaria, Spain  
Prof. Catarina Luísa Camarinhas, Universidade Técnica de Lisboa, Portugal  
Prof. Chris Stout, University of Illinois, IL, USA  
Prof. Dana Anderson, University of Colorado at Boulder, CO, USA  
Prof. Deolinda Rasteiro, Coimbra Institute of Engineering, Portugal  
Prof. Dmitrijs Serdjuks, Riga Technical University, Latvia  
Prof. Edy Portmann, University of Bern, Switzerland  
Prof. Eleazar Jimenez Serrano, Kyushu University, Japan  
Prof. F. G. Lupianez, University Complutense, Spain  
Prof. Fabio Nappo, University of Cassino and Southern Lazio, Italy  
Prof. Francesco Rotondo, Polytechnic of Bari University, Italy  
Prof. Francesco Zirilli, Sapienza Università di Roma, Italy  
Prof. Francisco Moya, University of Castilla-La Mancha, Spain  
Prof. Frederic Kuznik, National Institute of Applied Sciences, Lyon, France  
Prof. Garyfallos Arabatzis, University of Thrace, Greece  
Prof. Genqi Xu Tianjin, University, China  
Prof. George Barreto Pontificia, Universidad Javeriana, Colombia  
Prof. Guido Izuta, Yonezawa Women's College, Japan  
Prof. Guoxiang Liu, University of North Dakota, ND, USA  
Prof. Heimo Walter, Vienna University of Technology, Austria  
Prof. Hessam Ghasemnejad, Kingston University London, UK  
Prof. Hirofumi Nagashino, University of Tokushima, Japan  
Prof. Hongjun Liu, University of Notre Dame, IN, USA  
Prof. Hugo Rodrigues, Universidade Lusófona do Porto, Portugal  
Prof. Valeri Mladenov, Technical University of Sofia, Bulgaria  
Prof. James Vance, The University of Virginia's College at Wise, VA, USA  
Prof. João Bastos, Instituto Superior de Engenharia do Porto, Portugal  
Prof. John Cater, University of Auckland, New Zealand

Prof. José Carlos Metrôlho, Instituto Politecnico de Castelo Branco, Portugal  
Prof. Jose Flores, The University of South Dakota, SD, USA  
Prof. Kakuro Amasaka, Aoyama Gakuin University, Japan  
Prof. Karel Allegaert, University Hospitals Leuven, Belgium  
Prof. Kazuhiko Natori, Toho University, Japan  
Prof. Kei Eguchi, Fukuoka Institute of Technology, Japan  
Prof. Konstantin Volkov, Kingston University London, UK  
Prof. Kun Luo, Zhejiang University, China  
Prof. Kyandoghere Kyamakya, University of Klagenfurt, Austria  
Prof. Lapo Governi, University of Florence, Italy  
Prof. Lesley Farmer, California State University Long Beach, CA, USA  
Prof. Luigi Pomante, Università degli Studi dell'Aquila, Italy  
Prof. M. Javed Khan, Tuskegee University, AL, USA  
Prof. Maling Ebrahimpour, University of South Florida St Petersburg, FL, USA  
Prof. Manoj K. Jha, Morgan State University in Baltimore, USA  
Prof. Maria Ilaria Lunesu, University of Cagliari, Italy  
Prof. Mario Pestarino, University of Genova, Italy  
Prof. Masaji Tanaka, Okayama University of Science, Japan  
Prof. Mathieu Pétrissans, University of Lorraine, France  
Prof. Matteo Nunziati, University of Florence, Italy  
Prof. Matteo Palai, University of Florence, Italy  
Prof. Matthias Buyle, Artesis Hogeschool Antwerpen, Belgium  
Prof. Merzik Kamel, University of New Brunswick, Canada  
Prof. Miguel Carriegos, Universidad de Leon, Spain  
Prof. Minhui Yan, Shanghai Maritime University, China  
Prof. Mokhtari Fouad, University of Quebec at Trois-Rivières, Canada  
Prof. Moran Wang, Tsinghua University, China  
Prof. Najib Altawell, University of Dundee, UK  
Prof. Nicola Simola, University of Cagliari, Italy  
Prof. Nikola Vlahovic, University of Zagreb, Croatia  
Prof. Ole Christian Boe, Norwegian Military Academy, Norway  
Prof. Ottavia Corbi, University of Naples Federico II, Italy  
Prof. Pablo Fernandez de Arroyabe, University of Cantabria, Spain  
Prof. Pan Agathoklis, University of Victoria, Canada  
Prof. Pedro Lorca, University of Oviedo, Spain  
Prof. Philippe Dondon, Institut polytechnique de Bordeaux, France  
Prof. Philippe Fournier-Viger, University of Moncton, France  
Prof. Ricardo Gouveia Rodrigues, University of Beira Interior, Portugal  
Prof. Rocco Furferi, University of Florence, Italy  
Prof. Rosa Lombardi, University of Cassino and Southern Lazio, Italy  
Prof. Santoso Wibowo, CQ University, Australia  
Prof. Shinji Osada Gifu, University School of Medicine, Japan  
Prof. Sorinel Oprisan College of Charleston, SC, USA  
Prof. Stavros Ponis, National Technical University of Athens, Greece  
Prof. Sumanth Yenduri, University of Southern Mississippi, MS, USA  
Prof. Takuya Yamano, Kanagawa University, Japan  
Prof. Tetsuya Shimamura, Saitama University, Japan  
Prof. Tetsuya Yoshida, Hokkaido University, Japan  
Prof. Thomas Panagopoulos, University of Algarve, Portugal  
Prof. Tohru Kawabe, University of Tsukuba, Japan  
Prof. Vincenzo Niola, University of Naples Federico II, Italy  
Prof. Xiang Bai Huazhong, University of Science and Technology, China  
Prof. Xiaoguang Yue, Wuhan University of Technology, China  
Prof. Yamagishi Hiromitsu, Ehime University, Japan  
Prof. Yary Volpe, University of Florence, Italy

Prof. Yi Liang, Wuhan University, China  
Prof. Yuqing Zhou, Wuhan University of Technology, China  
Prof. Zhenbi Su, University of Colorado Boulder, CO, USA  
Prof. Zhong-Jie, Han Tianjin University, China

## Table of Contents

|   |    |
|---|----|
| <b>Keynote Lecture: Highlights of Modern Astrophysics, or The Gold Effect: How reliable is Modern Astrophysics?</b>                             | 14 |
| <i>Wolfgang Kundt</i>   |    |
| <b>Plenary Lecture: Data Compression: Learning and Clustering</b>   | 16 |
| <i>Bruno Carpentieri</i>  |    |
| <b>Silicon and CMOS-Compatible Spintronics</b>  | 17 |
| <i>Viktor Sverdlov, Joydeep Ghosh, Alexander Makarov, Thomas Windbacher, Siegfried Selberherr</i>   |    |
| <b>A Generalized Hebb (GH) Rule Based on a Cross-Entropy Error Function for Deep Belief Recursive Learning</b>                                  | 21 |
| <i>Mark J. Embrechts, Bernhard Sick</i>   |    |
| <b>Lossless Compression of Multidimensional Medical Images</b>  | 25 |
| <i>Raffaele Pizzolante, Bruno Carpentieri</i>   |    |
| <b>Does Time Pressure Induce Tunnel Vision? An Examination with the Eriksen Flanker Task by Applying the Hierarchical Drift Diffusion Model</b> | 30 |
| <i>Nico Assink, Rob H. J. Van Der Lubbe, Jean-Paul Fox</i>  |    |
| <b>Towards Community Recommendations on Location-Based Social Networks</b>  | 41 |
| <i>Chara Remoundou, Pavlos Kosmides, Konstantinos Demestichas, Ioannis Loumiotis, Evgenia Adamopoulou, Michael Theologou</i>                    |    |
| <b>Numerical Simulations of a Pipeline Crossing</b>   | 45 |
| <i>Ioan Both, Adrian Ivan</i>   |    |
| <b>Chromatics Aberrations of Diffractive Elements in Pulsed Laser Beams Formation</b>   | 50 |
| <i>Alexey P. Porfiriev, Sergey A. Degtyarev, Svetlana N. Khonina, Nikolai L. Kazanskiy</i>  |    |
| <b>Polarization Angle Independent Perfect Multi-Band Metamaterial Absorber in Microwave Frequency Regime</b>                                    | 54 |
| <i>O. T. Gunduz, C. Sabah</i>   |    |
| <b>The Simulation of Negative Influences in the Environment of Fixed Transmission Media</b>   | 58 |
| <i>Rastislav Róka</i>   |    |
| <b>Some Recent Advances of Ultrasonic Diagnostic Methods Applied to Materials and Structures (Including Biological Ones)</b>                    | 69 |
| <i>L. Nobile, S. Nobile</i>   |    |

|  |     |
|--|-----|
| <b>Mobility State Classification with Particle Filter</b>  | 75  |
| <i>Ha Yoon Song, Ji Hyun Baik</i>  |     |
| <b>The Impact of Memristive Devices and Systems on Nonlinear Circuit Theory</b>  | 83  |
| <i>Ricardo Riaza</i>   |     |
| <b>Detection Singular Polarization State by Multi-Order Diffractive Optical Element</b>  | 87  |
| <i>Dmitry A. Savelyev, Nikolay L. Kazanskiy, Svetlana N. Khonina</i>   |     |
| <b>The Structural Constant of an Atom as the Basis of Some Known Physical Constants</b>  | 92  |
| <i>Milan Perkovac</i>  |     |
| <b>Face-Recognition Based Authentication: Theory and Practice</b>  | 103 |
| <i>Thomas Fenzl, Christian Kollmitzer, Stefan Rass, Peter Schartner</i>  |     |
| <b>Co-Simulation of Redundant and Heterogeneous Modelling Scales for a Phenomenological Approach</b>   | 109 |
| <i>Sébastien Le Yaouanq, Christophe Le Gal, Pascal Redou, Jacques Tisseau</i>  |     |
| <b>Math Modelling of the Basic Defensive Activities</b>  | 116 |
| <i>Jan Mazal, Petr Stodola, Libor Kutěj, Milan Podhorec, Dana Křišťálová</i>   |     |
| <b>Polarization-Insensitive Perfect FSS Metamaterial Absorber in THz Frequency Range</b>   | 121 |
| <i>C. Sabah, F. Dincer, M. Karaaslan, E. Unal, O. Akgol</i>  |     |
| <b>The method of Probabilistic Nodes Combination in Simulation and Modeling</b>  | 124 |
| <i>Dariusz J. Jakóbczak</i>  |     |
| <b>Effect of Precursor on Growth of MoS<sub>2</sub> Monolayer and Multilayer</b>   | 130 |
| <i>Shraddha Ganorkar, Jungyoon Kim, Young Hwan Kim, Seong-Il Kim</i>   |     |
| <b>Photonic Crystal Cavities for Optical Signal Processing</b>   | 134 |
| <i>Nikolay L. Kazanskiy, Pavel G. Serafimovich</i>   |     |
| <b>A Stateless Key Management Technique for Protection of Sensitive Data at Proxy Level for SQL Based Databases Using NIST Recommended SP800-132</b>   | 140 |
| <i>Kurra Mallaiah, S. Ramachandram</i>   |     |
| <b>Time of flight Measurement Method to Determine the Milk Coagulation Cut Time</b>  | 147 |
| <i>Mourad Derra, Abdellah Amghar, Hassan Sahсах</i>  |     |
| <b>Ensurance and Simulation of Electromagnetic Compatibility: Recent Results in TUSUR University</b>   | 151 |
| <i>Talgat Gazizov, Alexander Melkozerov, Alexander Zabolotsky, Sergey Kuksenko, Pavel Orlov, Vasilij Salov, Roman Akhunov, Ilya Kalimulin, Roman Surovtsev, Maxim Komnatnov, Alexander Gazizov</i> |     |

|   |     |
|---|-----|
| <b>The First Principles Study on the TbP Compound</b>   | 163 |
| <i>Y. O. Ciftci, Y. Mogulkoc, M. Evecen</i>   |     |
| <b>Decision Making in Group Process of Consensus Based on Structures of Decision Dynamics: Application to the Superior Council of the UTEM</b>  | 173 |
| <i>Munoz S. Simon, Zapata C. Santiago</i>   |     |
| <b>Prediction of Fatigue Crack Propagation in Bonded Joints Using Fracture Mechanics</b>  | 186 |
| <i>Reza Hedayati, Meysam Jahanbakhshi</i>   |     |
| <b>Plastic Deformation and Fracture Processes in Layered Metal-Graphene Composites and Polycrystalline Graphene</b>                             | 193 |
| <i>Ilya A. Ovid'ko, Alexander G. Sheinerman</i>   |     |
| <b>Complex Social Network Interactions in Coupled Socio-Ecological System: Multiple Regime Shifts and Early Warning Detection</b>               | 196 |
| <i>Hendrik Santoso Sugiarto, Lock Yue Chew, Ning Ning Chung, Choy Heng Lai</i>  |     |
| <b>Progress in Ultrasonic Nano Manipulations</b>  | 205 |
| <i>Junhui Hu, Qiang Tang, Xu Wang, Xiaofei Wang</i>   |     |
| <b>The Gravity Control Experiments: Sensors, Equipment, Results</b>   | 210 |
| <i>Vitaly O. Groppen</i>  |     |
| <b>Influence of the Applied Electric Field on the Growth of an Electrical Discharge</b>   | 215 |
| <i>L. Zeghichi, L. Mokhnache, M. Djebabra</i>   |     |
| <b>A Novel Flexible Electrodynamical Planar Loudspeaker</b>   | 220 |
| <i>Jium-Ming Lin, Ubadigha Chinweze Ukachukwu, Cheng-Hung Lin</i>   |     |
| <b>Polarization Angle Independent Perfect Metamaterial Absorber</b>   | 225 |
| <i>C. Sabah, F. Dincer, E. Demirel, M. Karaaslan, E. Unal, O. Akgol</i>   |     |
| <b>On Adaptation Possibility of Model based on Slow Flow around Sphere for Determination of Flow Local Speeds in Window Between Spheres</b>     | 228 |
| <i>Anna Sandulyak, Darya Sandulyak, Olga Semina, Alexander Sandulyak</i>  |     |
| <b>Neural Networks Based Feature Selection from KDD Intrusion Detection Dataset</b>   | 232 |
| <i>Adel Ammar, Khaled Al-Shalfan</i>  |     |
| <b>Prospects of High-Frequency Gravimetry</b>   | 237 |
| <i>Alexander L. Dmitriev</i>  |     |
| <b>Variable Cosmological Parameter and S-channel Quantum Matter Fields Hadamard Renormalization in Spherically Symmetric Curved Space Times</b> | 241 |
| <i>Hossein Ghaffarnejad</i>   |     |
| <b>Authors Index</b>  | 249 |

## Keynote Lecture

### Highlights of Modern Astrophysics, or The Gold Effect: How reliable is Modern Astrophysics?



#### Professor Wolfgang Kundt

Argelander Institut für Astronomie, Bonn

GERMANY

E-mail: wkundt@astro.uni-bonn.de

**Abstract:** Our modern civilisation, on Earth, owes its existence and comfort to three kinds of machines:

(1) 'Manmade' machines, which produce (~millions of identical) cars, trains, airplanes, rockets, guns, bombs, smart phones, ipads, and of further useful equipment, which help us control our environment;

(2) 'Biological', or 'organic' machines, which produce (~millions of almost identical) living creatures, with all their senses for orientation, motion, action, and for housekeeping their bodies - steered by their DNA - and involving lenses, ears, thermostats, and all sorts of further organs acting like physical hardware; and:

(3) 'Inorganic' machines, which have been at work since the origin of the Universe - long before the existence of life - which provided stars, planets, moons, magnetic fields, accretion disks, stellar winds, nova and supernova explosions, and all the astrophysical twin-jets, also the cosmic rays, and the gamma-ray bursts. They are not always well understood.

Whereas all the manmade machines are well understood (by their constructors), and perform reliably, having been multiply tested, and all the organic machines often perform even better, and more reliably - having survived on Earth for millions, or even billions of years - the inorganic machines have often caused "conumbra" to their detectors, and have often found wrong explanations in the literature, suffering from the Gold effect: because they could not be tested. They were equally relevant for the existence of life. I will present various examples of the latter.

**Brief Biography of the Speaker:** Born in Hamburg in 1931; High School in Dresden and Hamburg; University Career in Hamburg: Diploma (1956), Ph.D. (1959), Habilitation (1965), all with Pasual Jordan in Theoretical Physics, centered on General Relativity.

Lectures in: Kiel, Hamburg, Geneva (CERN), Bielefeld, Bonn.

Extended Visits of: Pittsburgh (Pa), Cambridge (England), Kyoto, Bangalore, Boston, Linz, Maribor.

Organisation of 15 international Conferences in Europe.

Publication: of more than 280 articles, and 9 books on fundamental physics, including quantisation, astrophysics, geophysics, and biophysics, among them: "Astrophysics, A new Approach", Springer 2004, and "Physikalische Mythen auf dem Pruefstand", with Ole Marggraf, Springer 2014.

Scientific friends: Erich Bagge, Felix Pirani, Werner Israel, Thomas Gold, Hermann Bondi, Rolf Hagedorn, John Archibald Wheeler, Antonino Zichichi, Werner Buckel, Peter Scheuer, Rajaram Nityananda, David Layzer.

Frequent disagreements with mainstream opinions: More than 135 'alternatives' can be found in my publications, including my books.

## Plenary Lecture

### Data Compression: Learning and Clustering



**Professor Bruno Carpentieri**

Dipartimento di Informatica

Universita di Salerno

ITALY

E-mail: [bc@dia.unisa.it](mailto:bc@dia.unisa.it)

**Abstract:** Data Compression is generally motivated by the economic and logistic needs to save space in storage media and to save bandwidth in communication. Today we know that data compression, clustering, data classification, learning and data mining are all facets of the same multidimensional coin and that the data compression process is strictly bound to efficient clustering and learning. In this talk we will review some of the recent advances in the field and we will exploit the relationship between compression, learning and clustering.

**Brief Biography of the Speaker:** Bruno Carpentieri received the “Laurea” degree in Computer Science from the University of Salerno, Salerno, Italy, and the M.A. and Ph.D. degrees in Computer Science from the Brandeis University, Waltham, MA, U.S.A.

Since 1991, he has been first Assistant Professor and then Associate Professor of Computer Science at the University of Salerno (Italy). His research interests include lossless and lossy image compression, video compression and motion estimation, information hiding. He has been for many years Associate editor of the journal IEEE Trans. on Image Processing. He was chair and organizer of the International Conference on Data Compression, Communication and Processing 2011, co-chair of the International Conference on Compression and Complexity of Sequences, and, for many years, program committee member of the IEEE Data Compression Conference. He has been responsible for various European Commission contracts regarding image and video compression and digital movies.

# Silicon and CMOS-Compatible Spintronics

Viktor Sverdlov, Joydeep Ghosh, Alexander Makarov, Thomas Windbacher, and Siegfried Selberherr

Institute for Microelectronics, Technische Universität Wien

Gußhausstraße 27–29, A-1040 Wien, Austria

Email: {sverdlov|ghosh|makarov|windbacher|selberherr}@iue.tuwien.ac.at

**Abstract**—Silicon, the main material of microelectronics, is attractive for extending the functionality of MOSFETs by using the electron spin. However, spin lifetime decay in the silicon-on-insulator transistor channel is a potential threat to spin-driven devices using silicon. We predict a giant enhancement of the electron spin lifetime in silicon thin films by applying uniaxial mechanical stress. The advantage of our proposal is that stress techniques are routinely used in semiconductor industry to enhance the electron and hole mobilities in MOSFETs. The spin manipulation in silicon by purely electrical means is a challenge because of a relatively weak coupling of the electron spin to the electric field through an effective gate voltage dependent spin-orbit interaction. In contrast, the coupling between the spin orientation and charge current is much stronger in magnetic tunnel junctions. Magnetic random access memory (MRAM) built on magnetic tunnel junctions is CMOS-compatible and possesses all properties needed for universal memory. We demonstrate a significant improvement of one of the critical MRAM characteristics, the switching time, by specially designing the recording layer. Finally, by using MRAM arrays we discuss a realization of an intrinsic non-volatile logic-in-memory architecture suitable for future low-power electronics.

**Index Terms**—Spin lifetime modeling, valley splitting, tunneling magnetoresistance, magnetic tunnel junctions, spin transfer torque, universal memory, MRAM, implication-based logic, logic-in-memory

## I. INTRODUCTION

The tremendous increase in performance, speed, and density of modern integrated circuits has been supported by the continuous miniaturization of CMOS devices. Among the most crucial technological changes, lately adopted by the semiconductor industry, was the introduction of a new type of multi-gate three-dimensional (3D) transistors [1]. Multi-gate 3D device architecture potentially allows device scaling beyond 10nm. However, the obvious saturation of MOSFET miniaturization puts clear foreseeable limitations to the continuation of the increase in the performance of integrated circuits. Thus, research for finding alternative technologies and computational principles is paramount.

The electron spin is attractive for complimenting or even replacing the charge-based MOSFET functionality. It is characterized by two possible projections on a given axis and can be potentially used in digital information processing. In addition, only a small amount of energy is needed to alter the spin orientation. Silicon is well suited for spin-driven applications. Silicon predominantly consists of  $^{28}\text{Si}$  nuclei with zero magnetic spin. The spin-orbit interaction is also weak in the silicon conduction band. Even though these

features are promising, the demonstration of basic elements necessary for spin related applications, such as injection of spin-polarized currents into silicon, spin transport, and detection, was performed only recently.

A special technique [2] based on the attenuation of hot electrons with spins anti-parallel to the magnetization of the ferromagnetic film allowed creating an imbalance between the electrons with spin-up and spin-down in silicon thus injecting spin-polarized current. The spin-coherent transport through the device was studied by applying an external magnetic field causing precession of spins during their propagation from source to drain. The detection is performed with a similar hot electron spin filter. Although the drain current is fairly small due to the carriers' attenuation in the source and drain filters, as compared to the current of injected spins, the experimental set-up represents a first spin-driven device which can be envisaged working at room temperature. Contrary to the MOSFET, however, the described structure is a two-terminal device. Nevertheless, the first demonstration of coherent spin transport through an undoped  $350\mu\text{m}$  thick silicon wafer [3] has triggered a systematic study of spin transport properties in silicon [4].

## II. SILICON SPINFET

The spin field-effect transistor (SpinFET) is a future semiconductor spintronic device with a superior performance relative to the present transistor technology. Complementing or replacing the charge degree of freedom used for computation in modern CMOS circuits with the electron spin promises to reduce the energy dissipation [5]. SpinFETs are composed of two ferromagnetic contacts (source and drain), linked by a non-magnetic semiconductor channel region. The ferromagnetic contacts inject and detect spin-polarized electrons, in analogy to polarizer and analyzer as indicated already long ago by Datta and Das [6]. Because of the effective spin-orbit interaction in the channel, which depends on the perpendicular effective electric field, the spin of an electron injected from the source starts precessing. Only the electrons with spin aligned to the drain magnetization direction can leave the channel thus contributing to the current. Therefore, the total current through the device depends on the relative angle between the magnetization direction of the drain and the electron spin polarization at the end of the semiconductor channel. A current modulation is achieved by tuning the strength of the spin-orbit interaction in the semiconductor region by the gate voltage.. In order to realize the SpinFET, the following requirements must be fulfilled [7], namely an efficient spin injection and

This work is supported by the European Research Council through the grant #247056 MOSILSPIN.

detection, spin propagation, and spin manipulation by purely electrical means. We briefly overview recent achievements and challenges for the practical realization of a SpinFET.

### A. Spin injection and propagation

Spin injection in silicon from a ferromagnetic metal is overshadowed by an impedance mismatch problem [8]. A solution to the impedance mismatch problem is the introduction of a potential barrier between the ferromagnetic metal and the semiconductor [9]. A successful experimental demonstration of a signal which should correspond to spin injection in doped silicon at room temperature was first performed in 2009 [10] using an  $\text{Ni}_{80}\text{Fe}_{20}/\text{Al}_2\text{O}_3$  tunnel contact. Electrical spin injection through silicon dioxide at temperatures as high as 500K has been reported in [11]. Regardless of the success in demonstrating spin injection at room temperature, there are unsolved challenges which may compromise the results obtained. One of them is a several orders of magnitude discrepancy between the signal measured and the theoretical value. It turns out that the signal is stronger than predicted in three-terminal measurements [4]. The reasons for the discrepancies are heavily debated [12], [13] and it is apparent that more research is needed to resolve this controversy.

For a spin-based device the possibility to transfer the excess spin injected from the source to the drain electrode is essential. The excess spin is not a conserved quantity, in contrast to charge. While diffusing, it gradually relaxes to its equilibrium value which is zero in a non-magnetic semiconductor. An estimation for the spin lifetime at room temperature obtained within the three-terminal injection scheme was of the order 0.1-1ns [4]. This corresponds to an intrinsic spin diffusion length  $l=0.2\text{-}0.5\mu\text{m}$ . The spin lifetime is determined by the spin-flip processes. Several important spin relaxation mechanisms are identified [14], [15]. In silicon the spin relaxation due to the hyperfine interaction of spins with the magnetic moments of the  $^{29}\text{Si}$  nuclei is important at low temperature. Because of the inversion symmetry in the silicon lattice the Dyakonov-Perel spin relaxation mechanism is absent in bulk systems [14], [15]. At elevated temperatures the spin relaxation due to the Elliot-Yafet mechanism [14], [15] becomes important.

The Elliot-Yafet mechanism is mediated by the intrinsic interaction between the orbital motion of an electron and its spin and electron scattering. When the microscopic spin-orbit interaction is taken into account, the Bloch function with a fixed spin projection is not an eigenfunction of the total Hamiltonian. Because the eigenfunction contains a contribution with an opposite spin projection, even spin-independent scattering with phonons generates a small probability of spin flips [16]. The spin lifetime in undoped silicon at room temperature is about 10ns, which corresponds to a spin diffusion length of  $2\mu\text{m}$ , in agreement with experiments [4].

The main contribution to spin relaxation is due to optical phonon scattering between the valleys residing along different crystallographic axis, or  $f$ -phonons scattering [17], [18]. The relatively large spin relaxation reported in electrically-gated lateral-channel silicon structures [19], [20] indicates that the

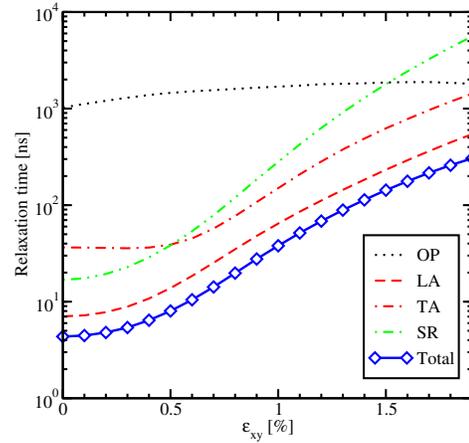


Fig. 1. Dependence of spin lifetime on shear strain for  $T=300\text{K}$  and a film of 4nm thickness. Optical (OP), longitudinal (LA) and transversal (TA) acoustic phonon, and surface roughness spin relaxation contributions are also shown.

extrinsic interface induced spin relaxation mechanism is important. This may pose an obstacle in realizing spin-driven CMOS-compatible devices, and a deeper understanding of the fundamental spin relaxation mechanisms in silicon inversion layers, thin films, and fins is needed.

The theory of spin relaxation must account for the most relevant scattering mechanisms which are due to electron-phonon interaction and surface roughness scattering. In order to evaluate the corresponding scattering matrix elements, the wave functions must be found. To find the wave functions, we employ the Hamiltonian describing the valley pairs along the [001]-axis [21]. The Hamiltonian includes confinement and an effective spin-orbit interaction. It also describes the unprimed subband structure, when uniaxial stress is applied [22]. Shear strain lifts the degeneracy between the unprimed subbands. The enhanced valley splitting makes the inter-valley spin relaxation irrelevant which results in a giant spin lifetime enhancement shown in Fig.1. Therefore, shear strain now routinely used to enhance the performance of modern MOSFETs can also be used to influence the spin propagation in the channel by enhancing the spin lifetime and the spin diffusion length significantly.

### B. Electric spin manipulation in silicon

Because of the weak spin-orbit interaction, silicon was not considered as a candidate for a SpinFET channel material. Recently it was shown [23] that thin silicon films inside SiGe/Si/SiGe structures may exhibit relatively large values of spin-orbit coupling. In actual samples with rough interfaces the inversion symmetry is broken, and atomistic simulations predict a relatively large value for the spin-orbit coupling  $\beta \approx 2\mu\text{eVnm}$  [24], which is in agreement with the only value reported experimentally [25], sufficient for realizing a silicon SpinFET. However, the channel length needed to achieve the substantial TMR modulation is close to a micron [26].

For shorter channels, the only option to realize a SpinFET is to exploit the relative magnetic orientation of the source and drain ferromagnetic contacts [7]. This adds a functionality to reprogram MOSFETs to obtain a different current under

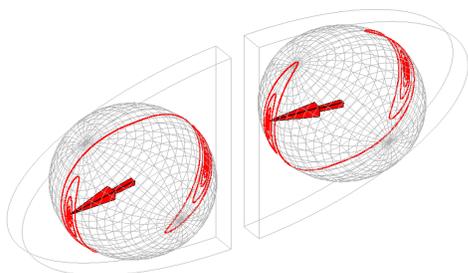


Fig. 2. Magnetization components vs. time for an elliptical  $52.5 \times 25\text{nm}^2$  MTJ with a composite free layer. The magnetizations of the left and right halves shown separately stay in-plane explaining the acceleration of switching (cf. [31]).

the same conditions by changing the drain magnetization orientation relative to the source. Once modified, the magnetization remains the same infinitely long without any external power. This property is used in emerging magnetic non-volatile memories.

### III. SPIN TRANSFER TORQUE MAGNETIC RAM

The basic element of magnetic random access memory (MRAM) is a magnetic tunnel junction (MTJ). The three-layer MTJ represents a sandwich of two magnetic layers separated by a thin spacer which forms a tunnel barrier. While the magnetization of the pinned layer is fixed, the magnetization orientation of the recording layer can be switched between the two stable states parallel and anti-parallel to the fixed magnetization direction. A memory cell based on MTJs is scalable, exhibits relatively low operating voltages, low power consumption, high operation speed, high endurance, and a simple structure.

Switching between the two states is induced by spin-polarized current flowing through the MTJ. The recording layer magnetization switching, by means of the spin transfer torque (STT) [27], [28], makes STT MRAM a promising candidate for future universal memory. Indeed, STT-MRAM is characterized by small cell size ( $4F^2$ ) and high density inherent to DRAM, fast access time (few ns) intrinsic to SRAM, non-volatility and long retention time subject to flash as well as high endurance ( $10^{14}$ ).

Because the spin-polarized current is only a fraction of the total charge current passing through the cell, high currents are required to switch the magnetization direction of the free layer. The reduction of the current density required for switching and the increase of the switching speed are the most important challenges in STT-MRAM developments [29]. If the recording layer is composed of two parts, one obtains a nearly three time faster switching at the same current density in a system with a composite in-plane ferromagnetic layer [30]. The composite layer is obtained by removing a central stripe of a certain width from the monolithic free layer of elliptic form (Fig.2). Due to the removal of the central region the switching processes of the left and right parts of the composite free layer occur in opposite senses to each other. In contrast to the switching of the monolithic layer, the magnetization of each half stays in-plane (Fig.2). This switching behavior leads to a decrease

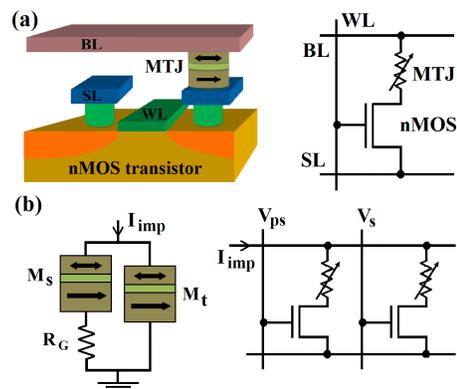


Fig. 3. The common STT-MRAM architecture based on the one-transistor/one-MTJ (1T/1MTJ) structure (cf. [35]).

of the switching energy barrier, while preserving the thermal stability. The reduction of the switching time depending on geometry parameters is investigated in [31], [32].

### IV. STT-MRAM BASED LOGIC-IN-MEMORY

The introduction of non-volatile logic could help significantly reducing the heat generation, especially at stand-by, booting, and resuming stages. It is extremely attractive to use the same elements as memory and latches to reduce the time delay and energy waste while transferring data between CPU and memory blocks. STT-MTJ-based memory has all the characteristics of a universal memory [36]. Furthermore, the MTJ technology is attractive for building logic configurations which combine non-volatile memory cells and logic circuits (so-called logic-in-memory architecture) to overcome the leakage power issue [37]–[39].

Recently, the realization of MTJ-based non-volatile logic gates was successfully demonstrated, for which the MTJ devices are used simultaneously as non-volatile memory cells and main computing elements [33]–[35]. In [33], [34] re-programmable logic gates realize the basic Boolean logic operations AND, OR, NAND, NOR, and the Majority operation. All basic Boolean logic operations are executed in two sequential steps including an appropriate preset operation in the output MTJ and then applying a voltage pulse ( $V_A$ ) with a proper amplitude to the gate. Depending on the logic states of the input MTJs, the preset in the output MTJ, and the voltage level applied to the gate, a conditional switching behavior in the output MTJ is provided, which corresponds to a particular logic operation [34]. A different set of the MTJ-based logic gates [35] is designed by using any two MRAM memory cells from an array to realize the Boolean implication (IMP) operation (Fig.3). Compared to the  $\text{TiO}_2$  memristive switches [40], MTJs provide a higher endurance. Furthermore, the bistable resistance state of the MTJs eliminates the need for refreshing circuits. The logic implementation using MTJ-based gates relies on a conditional switching provided by the state-dependent current modulation on the output (target) MTJ. The resistance modulation between the high and low resistance states in the MTJ is proportional to the TMR ratio. The error probability of MTJ-based operations decrease with increasing

TMR ratio which is thus the most important device parameter for the reliability [41].

## V. SUMMARY AND CONCLUSION

Recent ground-breaking experimental and theoretical findings regarding spin injection and transport in silicon make spin an attractive option to supplement or to replace the charge degree of freedom for computations. Stress routinely used to enhance the electron mobility can also be used to boost the spin lifetime. CMOS-compatible STT-MRAM cells built on magnetic tunnel junctions with a composite recording layer demonstrate a three-fold improvement of the switching time as compared to similar cells with a monolithic layer. The realization of an intrinsic non-volatile logic-in-memory architecture by using MRAM arrays is outlined.

## REFERENCES

- [1] M. Bohr, "The evolution of scaling from the homogeneous era to the heterogeneous era," in *International Electron Devices Meeting (IEDM)*, 2011, pp. 1.1.1–1.1.6.
- [2] I. Appelbaum, B. Huang, and D. J. Monsma, "Electronic measurement and control of spin transport in silicon," *Nature*, vol. 447, pp. 295–298, 2007.
- [3] B. Huang, D. J. Monsma, and I. Appelbaum, "Coherent spin transport through a 350 micron thick silicon wafer," *Phys. Rev. Lett.*, vol. 99, p. 177209, Oct 2007.
- [4] R. Jansen, "Silicon spintronics," *Nature Materials*, vol. 11, pp. 400–408, 2012.
- [5] S. Bandyopadhyay and M. Cahay, "Electron spin for classical information processing: A brief survey of spin-based logic devices, gates and circuits," *Nanotechnology*, vol. 20, no. 41, p. 412001, 2009.
- [6] S. Datta and B. Das, "Electronic analog of the electro-optic modulator," *Appl. Phys. Lett.*, vol. 56, no. 7, pp. 665–667, 1990.
- [7] S. Sugahara and J. Nitta, "Spin-transistor electronics: An overview and outlook," *Proc. IEEE*, vol. 98, no. 12, pp. 2124–2154, Dec 2010.
- [8] G. Schmidt, D. Ferrand, L. W. Molenkamp, A. T. Filip, and B. J. van Wees, "Fundamental obstacle for electrical spin injection from a ferromagnetic metal into a diffusive semiconductor," *Phys. Rev. B*, vol. 62, pp. R4790–R4793, Aug 2000.
- [9] E. I. Rashba, "Theory of electrical spin injection: Tunnel contacts as a solution of the conductivity mismatch problem," *Phys. Rev. B*, vol. 62, pp. R16267–R16270, Dec 2000.
- [10] S. P. Dash, S. Sharma, R. S. Patel, M. P. de Jong, and R. Jansen, "Electrical creation of spin polarization in silicon at room temperature," *Nature*, vol. 462, pp. 491–494, 2009.
- [11] C. Li, O. van't Erve, and B. Jonker, "Electrical injection and detection of spin accumulation in silicon at 500K with magnetic metal/silicon dioxide contacts," *Nature Communications*, vol. 2, p. 245, 2011.
- [12] R. Jansen, A. M. Deac, H. Saito, and S. Yuasa, "Injection and detection of spin in a semiconductor by tunneling via interface states," *Phys. Rev. B*, vol. 85, p. 134420, Apr 2012.
- [13] Y. Song and H. Dery, "Magnetic-field-modulated resonant tunneling in ferromagnetic-insulator-nonmagnetic junctions," *Phys. Rev. Lett.*, vol. 113, p. 047205, Jul 2014.
- [14] I. Zutic, J. Fabian, and S. Das Sarma, "Spintronics: Fundamentals and applications," *Rev. Mod. Phys.*, vol. 76, pp. 323–410, Apr 2004.
- [15] J. Fabian, A. Matos-Abiaguea, C. Ertler, P. Stano, and I. Zutic, "Semiconductor spintronics," *Acta Phys. Slovaca*, vol. 57, pp. 565–907, 2007.
- [16] J. L. Cheng, M. W. Wu, and J. Fabian, "Theory of the spin relaxation of conduction electrons in silicon," *Phys. Rev. Lett.*, vol. 104, p. 016601, Jan 2010.
- [17] P. Li and H. Dery, "Spin-orbit symmetries of conduction electrons in silicon," *Phys. Rev. Lett.*, vol. 107, p. 107203, Sep 2011.
- [18] Y. Song and H. Dery, "Analysis of phonon-induced spin relaxation processes in silicon," *Phys. Rev. B*, vol. 86, p. 085201, Aug 2012.
- [19] J. Li and I. Appelbaum, "Modeling spin transport in electrostatically-gated lateral-channel silicon devices: Role of interfacial spin relaxation," *Phys. Rev. B*, vol. 84, p. 165318, Oct 2011.
- [20] —, "Lateral spin transport through bulk silicon," *Appl. Phys. Lett.*, vol. 100, no. 16, pp. 162408, 2012.
- [21] D. Osintsev, O. Baumgartner, Z. Stanojevic, V. Sverdlov, and S. Selberherr, "Subband splitting and surface roughness induced spin relaxation in (001) silicon SOI MOSFETs," *Solid-State Electron.*, vol. 90, pp. 34 – 38, 2013.
- [22] V. Sverdlov, *Strain-Induced Effects in Advanced MOSFETs*. Wien - New York: Springer, 2011.
- [23] J.-M. Jancu, J.-C. Girard, M. O. Nestoklon, A. Lemaître, F. Glas, Z. Z. Wang, and P. Voisin, "STM images of subsurface Mn atoms in GaFs: Evidence of hybridization of surface and impurity states," *Phys. Rev. Lett.*, vol. 101, p. 196801, Nov 2008.
- [24] M. Prada, G. Klimeck, and R. Joynt, "Spin-orbit splittings in Si/SiGe quantum wells: From ideal Si membranes to realistic heterostructures," *New Journal of Physics*, vol. 13, no. 1, p. 013009, 2011.
- [25] Z. Wilamowski and W. Jantsch, "Suppression of spin relaxation of conduction electrons by cyclotron motion," *Phys. Rev. B*, vol. 69, p. 035328, Jan 2004.
- [26] D. Osintsev, V. Sverdlov, Z. Stanojević, A. Makarov, and S. Selberherr, "Temperature dependence of the transport properties of spin field-effect transistors built with InAs and Si channels," *Solid-State Electron.*, vol. 71, pp. 25 – 29, 2012.
- [27] J. Slonczewski, "Current-driven excitation of magnetic multilayers," *Journal of Magnetism and Magn. Materials*, vol. 159, no. 1-2, pp. L1–L7, 1996.
- [28] L. Berger, "Emission of spin waves by a magnetic multilayer traversed by a current," *Phys. Rev. B*, vol. 54, pp. 9353–9358, Oct 1996.
- [29] A. V. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepurskii, R. S. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. H. Butler, P. B. Visscher, D. Lottis, E. Chen, V. Nikitin, and M. Krounbi, "Basic principles of STT-MRAM cell operation in memory arrays," *J. Phys. D*, vol. 46, no. 7, p. 074001, 2013.
- [30] A. Makarov, V. Sverdlov, D. Osintsev, and S. Selberherr, "Reduction of switching time in pentalayer magnetic tunnel junctions with a composite-free layer," *Phys. Status Solidi Rapid Research Letters*, vol. 5, no. 12, pp. 420–422, 2011.
- [31] A. Makarov, V. Sverdlov, and S. Selberherr, "Magnetic tunnel junctions with a composite free layer: A new concept for future universal memory," in *Future Trends in Microelectronics*. John Wiley & Sons, 2013, pp. 93–101.
- [32] A. Makarov, "Modeling of emerging resistive switching based memory cells," Dissertation, Institute for Microelectronics, TU Wien, 2014.
- [33] A. Lyle, J. Harms, S. Patil, X. Yao, D. J. Lilja, and J.-P. Wang, "Direct communication between magnetic tunnel junctions for nonvolatile logic fan-out architecture," *Appl. Phys. Lett.*, vol. 97, no. 15, p. 152504, 2010.
- [34] A. Lyle, S. Patil, J. Harms, B. Glass, X. Yao, D. Lilja, and J.-P. Wang, "Magnetic tunnel junction logic architecture for realization of simultaneous computation and communication," *IEEE Trans. Magn.*, vol. 47, no. 10, pp. 2970–2973, Oct 2011.
- [35] H. Mahmoudi, T. Windbacher, V. Sverdlov, and S. Selberherr, "Implication logic gates using spin-transfer-torque-operated magnetic tunnel junctions for intrinsic logic-in-memory," *Solid-State Electron.*, vol. 84, pp. 191 – 197, 2013.
- [36] C. Augustine, N. Mojumder, X. Fong, H. Choday, S. P. Park, and K. Roy, "STT-MRAMs for future universal memories: Perspective and prospective," in *International Conference on Microelectronics (MIEL)*, May 2012, pp. 349–355.
- [37] T. Endoh, "STT-MRAM technology and its NV-logic applications for ultimate power management," in *CMOS Emerging Technologies Research (CMOSETR)*, 2014, p. 14.
- [38] W. Zhao, E. Belhaire, C. Chappert, F. Jacquet, and P. Mazoyer, "New non-volatile logic based on spin-MTJ," *Phys. Status Solidi A*, vol. 205, no. 6, pp. 1373–1377, 2008.
- [39] M. Natsui, D. Suzuki, N. Sakimura, R. Nebashi, Y. Tsuji, A. Morioka, T. Sugibayashi, S. Miura, H. Honjo, K. Kinoshita, S. Ikeda, T. Endoh, H. Ohno, and T. Hanyu, "Nonvolatile logic-in-memory array processor in 90nm MTJ/MOS achieving 75% leakage reduction using cycle-based power gating," in *International Solid-State Circuits Conference (ISSCC)*, Feb 2013, pp. 194–195.
- [40] J. Borghetti, G. Snider, P. Kuekes, J. Yang, D. Stewart, and R. Williams, "Memristive switches enable stateful logic operations via material implication," *Nature*, vol. 464, pp. 873–876, 2010.
- [41] H. Mahmoudi, T. Windbacher, V. Sverdlov, and S. Selberherr, "Reliability analysis and comparison of implication and reprogrammable logic gates in magnetic tunnel junction logic circuits," *IEEE Trans. Magn.*, vol. 49, no. 12, pp. 5620–5628, Dec 2013.

# A Generalized Hebb (GH) rule based on a cross-entropy error function for deep belief recursive learning

Mark J. Embrechts and Bernhard Sick

**Abstract**— The purpose of this article is to show a novel learning rule derived from the Hebb rule, for the training of recursive deep belief networks. The derivation is de novo, elegant, and similar to Hinton’s Contrastive Divergence CD-N, where N is the number of recursions (e.g., 1 in a traditional neural multi-layered perceptron). This rule is introduced in this article as “*Generalized Hebb rule*” (GH-N) for an entropic cost function for deep belief recursive learning. This rule is important because: (i) It is easy to apply, (ii) it applies to a stacked deep belief network, (iii) Hinton’s Contrastive Divergence rule CD-N for continuous units is a special case of this rule, and (iv) preliminary experimental results show that – for binary patterns – a deep belief auto-associator trained with a recursive neural network often shows a clearer separation of classes in the bottleneck layer than trained with backpropagation or compared to principal component analysis.

**Keywords**—deep belief network, Generalized Hebb rule, contrastive divergence, auto-associator.

## I. MOTIVATION

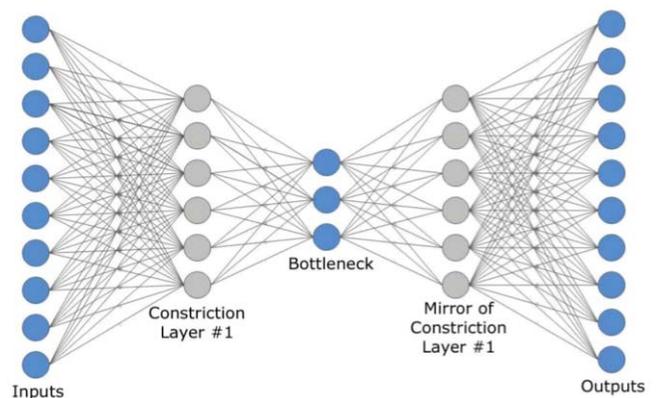
DEEP belief networks (DBN) [1-2] have rejuvenated interest in artificial neural networks, but are still hard to grasp for novices in artificial neural networks. DBN are basically a stacking of layers of neurons and can be trained layer by layer using Restricted Boltzmann Machines (RBM) [3-6]. The purpose of this article is to introduce and derive a novel training rule, the *Generalized Hebb rule (GH-N)*, for recursive deep belief stacked auto-associators, which can also be applied to deep belief networks. The resulting training rule has a strong similarity with Hinton’s Contrastive Divergence rule (CD-N) [6-8] but applies directly to continuous units [16] as well, and does not need the Boltzmann type of “stochasticity” to interpret the firing of a neuron. This rule is derived de novo, starting from the Hebb rule, and applies to a recursive single layer of a stacked auto-associator. Preliminary tests using the Italian olive oil data [11-12] show that the bottleneck neuron outputs of a deep belief recursive stacked auto-associator for binary

and multi-class classification patterns show a clearer separation on the test data than a deep belief network trained with backpropagation based on LeCun’s Efficient BackProp [9-10].

This article is organized as follows: Section II shows an intuitive derivation of the GH-N algorithm; Section III addresses how this rule can be implemented for a stacked auto-associator; Section IV discusses preliminary experiments; and Section V summarizes the key findings and gives an outlook to future work.

## II. DEEP BELIEF AUTO-ASSOCIATIVE NEURAL NETWORKS

A deep belief auto-associative neural network is an auto-associator [13-15] with many layers, usually with symmetric weights, and trained with a deep belief method. An auto-associator can be regarded as an artificial neural network, where the output values (i.e., the target values) are exactly the same as the input values. Such an auto-associator has many layers of neurons and a bottleneck layer.



**Figure 1.** Example of a deep auto-associator where the target outputs are the same as the inputs. Except for the input layer, all other elements represented by a circle are artificial neurons. It can be shown that the weights are symmetric. [21]

The auto-associator depicted in Fig. 1 has a symmetric structure and in this case, also symmetric weights. Usually the bottleneck layer is often of special interest and can be used in a similar manner as principal components and/or independent components. Fig. 1 represents the scheme of a deep auto-associator, where the outputs are the same as the inputs. Ex-

This work was supported in part by the German Federal Ministry of Education and Research under Grant 03EK3536A (PrIME project).

M. J. Embrechts is with CardioMag Imaging, Inc., 13 British American Boulevard, Latham, NY 12110, U.S.A. ([mark.embrechts@gmail.com](mailto:mark.embrechts@gmail.com)).

B. Sick is with the University of Kassel, Germany, Electrical Engineering and Computer Science Department, Intelligent Embedded Systems Lab, Wilhelmshoher Allee 71-73, 34121 Kassel, Germany ([bsick@uni-kassel.de](mailto:bsick@uni-kassel.de)).

cept for the input layer, all other elements represented by a circle are artificial neurons (i.e., first computing a weighted sum of the inputs, and then applying a nonlinear activation function: typically a sigmoid or a hyperbolic tangent function).

### III. DERIVATION OF THE GENERALIZED HEBB RULE

It is in principle possible to train an auto-associative network via the backpropagation algorithm [17], and a deep auto-associative network via Efficient BackProp [9-10]. However, a stepwise building up of a deep belief auto-associative network [10] can easier avoid that neurons get into saturation and possibly reduce the training time, too. A stepwise deep belief auto-associator can be trained – layer by layer – by Hinton’s Contrastive Divergence rule CD-N [7-8], or by our new Generalized Hebb rule, GH-N.

The derivation of the Generalized Hebb rule will proceed as follows: (i) First we will derive the delta rule as an extension of the Hebb rule; (ii) using this rule we will derive the training rule for a single layer of a symmetric recursive auto-associator; (iii) then we will derive the training rule of a stacked symmetric auto-associator. A comparison of the GH-N rule with Hinton’s CD-N will then be made.

#### A. A de novo derivation of the Widrow-Hoff Delta rule from Hebb’s rule

The Widrow-Hoff Delta rule can be considered as an extension of Hebb’s rule, which states: “When an axon of a cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased.” [10, 18].

The above rule is a “carrot rule”: i.e., good behavior gets rewarded. We can extend this to “a carrot and a stick” rule: i.e., good behavior gets rewarded and bad behavior gets punished.

Rosenblatt’s single-layer Perceptron rule for binary patterns can be derived from that principle and written more elegantly for polar (i.e., [-1,1]) units. In this case, the output of a neural network is the weighted sum of the inputs (modified by a bias  $b$ ) according to:

$$\text{output} = \sum_{i=1}^m w_i x_i + b.$$

Patterns are shown one-at-a-time. If a pattern is classified correctly, the weights are not modified. If a pattern is misclassified, one of the following rules is applied:

$$\vec{w}^{(N+1)} = \vec{w}^{(N)} - \eta^{(N)} \vec{x}^{(N)}$$

$$\vec{w}^{(N+1)} = \vec{w}^{(N)} + \eta^{(N)} \vec{x}^{(N)},$$

depending on whether  $x$  belongs to the negative or the positive class [10].  $N$  indicates the update iteration level.

For continuous units and supervised learning both rules can be combined into the Widrow-Hoff delta rule

$$\vec{w}_{ji}^{(N+1)} = w_{ji}^{(N)} - \eta \delta_j \vec{x}_i,$$

where delta is the error. Here, we use the following notation:  $w_{ji}$  is a weight for the connection from neuron  $i$  on the input side to neuron  $j$  on the output side.

In the backpropagation algorithm  $\delta'$  is used rather than  $\delta$ , where  $\delta' = (y_j - x_j) f'(x_j)$ . In case just  $\delta$  is used, this is also equivalent to a backpropagation rule where a different cost function than the least-squares error is applied, the bi-level entropic error function described by Baum [19,20]:

$$C(x, y) = - \sum_{\mu} [x^{\mu} \log(y^{\mu}) + (1 - x^{\mu}) \log(1 - y^{\mu})].$$

#### B. Stacked (recursive) symmetric auto-associator

A stacked auto-associator with symmetric weights is shown in the left hand side of the figure below. In the unfolded recursive auto-encoder with shared weights the weights are not shown. For symmetric weights and three recursions in the auto-encoder the Widrow-Hoff delta rule with a bi-level entropic error function and three recursions the learning rule is shown below.

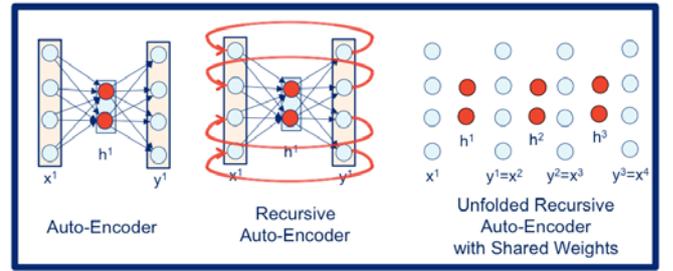


Figure 2. Recursive auto-associator with cross-entropic cost function, symmetric weights, weight sharing and delta rule through time approach with a 3-step example.

Note that we used the notation as explained in the right hand-side of Fig. 2, where the auto-encoder with three recursions is unfolded. We now use symmetric weights (i.e., only the weights of the second layer in the auto-encoder are trained, the weights in the first layer are just copied, using the weight symmetry property). Note also that we use weight sharing, which explains why the respective weight updates contain the factor  $1/3$ .

$$\Delta w_{ji} = \frac{\eta}{3} \langle [h_i^{(1)}(x_j^{(1)} - y_j^{(1)})] + [h_i^{(2)}(x_j^{(2)} - y_j^{(2)})] + [h_i^{(3)}(x_j^{(3)} - y_j^{(3)})] \rangle$$

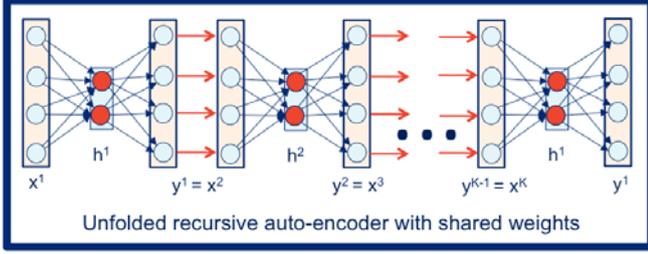
$$\Delta w_{ji} = \frac{\eta}{3} \langle [h_i^{(1)}(x_j^{(1)} - x_j^{(2)})] + [h_i^{(2)}(x_j^{(2)} - x_j^{(3)})] + [h_i^{(3)}(x_j^{(3)} - y_j^{(3)})] \rangle$$

$$\Delta w_{ji} = \frac{\eta}{3} \langle [h_i^{(1)}x_j^{(1)} + [x_j^{(2)}(h_j^{(2)} - h_j^{(1)})] + [x_j^{(3)}(h_j^{(3)} - h_j^{(2)})] - h_i^{(3)}y_j^{(3)} \rangle$$

$$\Delta w_{ji} \cong \eta' \langle [x_j^{(1)}h_i^{(1)} - y_j^{(3)}h_i^{(3)}] \rangle$$

### C. Stacking several recursive symmetric auto-associators

Fig. 3 expands to concept of Fig. 2 to an indefinite level of  $K$  recursions.



**Figure 3.** Recursive auto-associator with cross-entropy cost function, symmetric weights, weight sharing and delta rule through time approach (general  $\rightarrow K$  steps).

The learning formula can now be approximated as shown below. An explanation of the bracket notation is now in order. The bracket notation shows that all the patterns need to be applied. Note that we assumed in the notation a more or less convergent behavior from one recursive auto-associator to the next layer.

$$\begin{aligned} \Delta w_{ji} &= \frac{\eta}{K} \left[ \langle h_j^{(1)} (x_j^{(1)} - y_j^{(1)}) \rangle + \langle h_j^{(2)} (x_j^{(2)} - y_j^{(2)}) \rangle + \dots + \langle h_j^{(K)} (x_j^{(K)} - y_j^{(K)}) \rangle \right] \\ \Delta w_{ji} &= \frac{\eta}{K} \left[ \langle h_j^{(1)} (x_j^{(1)} - x_j^{(2)}) \rangle + \langle h_j^{(2)} (x_j^{(2)} - x_j^{(3)}) \rangle + \dots + \langle h_j^{(3)} (x_j^{(3)} - y_j^{(K)}) \rangle \right] \\ \Delta w_{ji} &= \frac{\eta}{K} \left[ \langle h_j^{(1)} x_j^{(1)} \rangle + \langle x_j^{(2)} (h_j^{(2)} - h_j^{(1)}) \rangle + \dots + \langle x_j^{(K)} (h_j^{(K)} - h_j^{(K-1)}) \rangle - h_j^{(K)} y_j^{(K)} \right] \\ \Delta w_{ji} &\cong \eta \left[ \langle x_j^{(1)} h_j^{(1)} \rangle - \langle y_j^{(K)} h_j^{(K)} \rangle \right] \\ \Delta w_{ji} &\cong \eta \left[ \langle x_j^{(1)} h_j^{(1)} \rangle - \langle y_j^{(K)} h_j^{(K)} \rangle \right] \end{aligned}$$

### D. The Generalized Hebb rule GH-N for deep belief auto-encoders

In short, the Generalized Hebb rule for updating the weight  $w_{ji}$  can be written as:

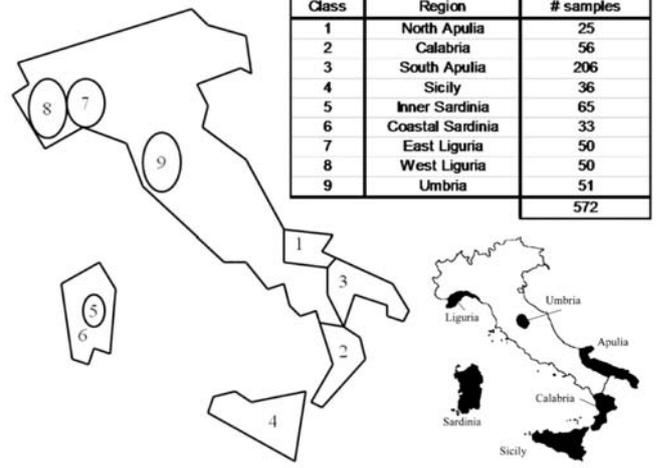
$$\begin{aligned} \Delta w_{ji}^{GH-K} &= \eta \cdot \\ &\left[ \langle x_j^{(1)} h_i^{(1)} \rangle + \sum_{k=2}^{K-1} \langle x_j^{(k)} (h_j^{(k)} - h_j^{(k-1)}) \rangle - \langle y_j^{(K)} h_i^{(K)} \rangle \right] \end{aligned}$$

Assuming that we are near a converging behavior for the output of the hidden layer, this rule can be approximated by the well-known Contrastive Divergence rule for continuous (non-stochastic) units:

$$\Delta w_{ji}^{CD-K} = \eta \cdot \left[ \langle x_j^{(1)} h_i^{(1)} \rangle - \langle y_j^{(K)} h_i^{(K)} \rangle \right].$$

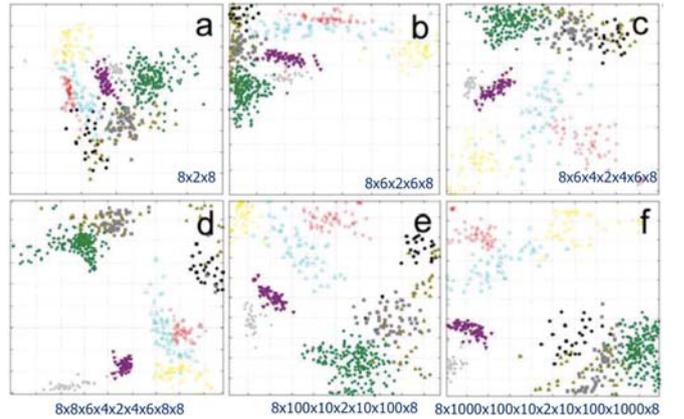
## IV. PRELIMINARY RESULTS

We will illustrate this procedure on the Italian olive oil data [11-12]. In this case there are 572 olive oils (Fig. 4) from different regions in Italy, described by 8 different fatty acids.



**Figure 4.** Presentation of 572 Italian olive oils. The olive oils are described by measures for 8 different fatty acids. Note that the 9 classes of olive oils are not balanced.

Fig. 5 shows several deep belief network results for the Italian olive oil data. Fig. 5a is equivalent to a principal component projection on the first two principal components, while 5b – 5c are results from a backpropagation algorithm for deep belief networks for different network structures. Even though in this particular case the results were obtained from applying the backpropagation algorithm without recursion, the GH-1 results are of a similar nature and also show a much clearer separation than the principal components. Note that the northern and southern Italian olive oils become more clearly separated the deeper the network is.



**Figure 5.** Projections in the bottleneck layer of 572 Italian olive oil data for various deep belief neural network structures. The olive oils become more clearly separated, the deeper the network structure is (cf. [21]).

## V. CONCLUSION

This paper introduced a novel Generalized Hebb rule (GH-N) as an alternate to Hinton's Contrastive Divergence rule (CD-N) for training deep belief networks. While both rules have many similarities, the emphasis of this paper is on a simple derivation from basic principles.

In our future work we will investigate the theoretical behavior and the actual performance of the novel GH-N technique in much more detail.

## REFERENCES

- [1] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin [2009] Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, Vol. 1, pp. 1-40.
- [2] Yoshua Bengio [2012] *Learning Deep Architectures for AI*. Technical Report 1312, University of Monreal, Canada.
- [3] Geoffrey E. Hinton, and Terrence J. Sejnowski [1986] Learning and relearning in Boltzmann machines. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press. Cambridge, MA, Vol. 1, pp. 283-317.
- [4] Hsin Chen and Alan. F. Murray [2003] Continuous restricted Boltzmann machine with an implementable training algorithm. *IEEE Proceedings of Visual Image and Signal Processing*, Vol. 150(3), pp. 153-158.
- [5] Benjamin M. Marlin, Kevin Swersky, Bo Chen, and Nando de Freitas [2010] Inductive principles for restrictive Boltzmann Machine Learning. *Proceedings of the 13<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS)*, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9, pp. 509-516.
- [6] Geoffrey Hinton [2010] A practical guide to training Restricted Boltzmann Machines, Version 1. *University of Toronto Technical Report*, UTML TR 2010-003 [Augst2, 2010].
- [7] Ilya Sutskever and Tijmen Tieleman [2010] On the convergence properties of Contrastive Divergence. *Proceedings of the 13<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS)*. May 13-15, Sardinia, Italy.
- [8] Geoffrey Hinton [2002] Training products of experts by minimizing contrastive divergence. *Neural Computation*, Vol. 14(8), pp. 1771-1800.
- [9] Yan LeCun, L. Bottou, G. Orr and K. Muller [1988] Efficient BackProp. In Orr, G. and Muller, K. (Eds.) *Neural Networks: Tricks of the Trade*. Springer
- [10] Simon Haykin [2009] *Neural Networks and Learning Machines, Third Edition*, Pearson.
- [11] Michele Forina and Carla Armanino [1981] Eigenvector projection and simplified nonlinear mapping of fatty acid content of Italian olive oils. *Ann. Chem.*, Vol. 72, pp. 125-127.
- [12] Jure Zupan and Johann Gasteiger [1999] *Neural Networks in Chemistry and Drug Design (2<sup>nd</sup> Edition)*. Wiley-VCH.
- [13] Hervé Bourlard and Yves Kamp, Auto-association by multilayer perceptrons and singular value decomposition, *Biological Cybernetics*, Vol. 59, pp. 291-294, 1988.
- [14] Nathalie Japkowicz, Stephen J. Hanson, and Mark A. Gluck, Nonlinear autoassociation is not equivalent to PCA, *Neural Computation*, Vol. 12, pp. 531-545, MIT, 2000.
- [15] M. A. Kramer, Autoassociative neural networks, *Computers and Chemical Engineering*, Vol.16, pp. 313-328, Pergamon Press, 1992.
- [16] Hugo LaRochelle, Benjamin Bengio, Jérôme Louradour, and Pascal Lamblin [2007] Exploring strategies for training deep belief networks. *Journal of Machine Learning Research*, Vol. 1, pp. 1-40.
- [17] Paul J. Werbos [1994] *The Roots of Backpropagation. From Ordered Derivatives to Neural Networks and Political Forecasting*. New York, NY. John Wiley & Sons, Inc.
- [18] Donald O. Hebb [1949] *The Organization of Behavior*. New York, Wiley and Sons.
- [19] Eric B. Baum, and Frank Wilzek [1988] Supervised learning of probability distributions by neural networks. In *Neural Information Processing Systems*. Denver 1987. D. Z. Anderson, Editor, pp. 52-56.
- [20] Pascal Vincent, Hugo LaRochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol [2010] Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, Vol. 11, pp. 3371-3408.
- [21] Mark J. Embrechts, Blake Hargis, and Jonathon D. Linton, "Augmented Efficient BackProp for Backpropagation Learning in Deep Autoassociative Neural Networks." *Proceedings of the 2010 IEEE International Joint Conference on Neural Networks (IJCNN 2010)* as part of The World Congress on Computational Intelligence (WCCI 2010), pp. 1012-1020, Barcelona, Spain, July 18-23, 2010.

# Lossless Compression of Multidimensional Medical Images

Raffaele Pizzolante, Bruno Carpentieri

**Abstract**— We introduce an efficient lossless algorithm that can be used for the compression of multidimensional medical images. We experimentally test our approach on a test set of 3-D computed tomography (CT) and 3-D magnetic resonance (MR) images. The achieved results outperform other state-of-the-art approaches.

**Keywords**— multidimensional medical images compression; multidimensional medical images coding; multidimensional data compression;

## I. INTRODUCTION

TODAY medical digital imaging techniques are continuously evolving. The research activities in medical imaging focus primarily on the improvement of the acquisition and transmission algorithms. Thanks to the wide diffusion of inter-connections new services are provided to medical staffs: for examples, exchange of medical data among different entities/structures connected by networks (e.g. through Internet, Clouds services, P2P networks, etc.), telemedicine, tele-radiology, real-time tele-consultation, PACS (Picture Archiving and Communication Systems), etc..

For all these applications one of the main disadvantages is related to the large amount of storage space needed to save the images and for the time required to transmit the data.

These costs grow proportionally to the size of data. Future expectations in medical applications will further increase the requests for memory space and/or efficient transmission time.

Different medical imaging methodologies produce multidimensional data. For instance, Computed Tomography (CT) and Magnetic Resonance (MR) imaging technologies produce three-dimensional ( $N=3$ ) data.

A 3-D CT image is acquired through X-rays. The acquisition process is performed via a computer. By using the computer we are able to obtain different cross-sectional views.

3-D CT images are an important tool for the identification of normal or abnormal structures of the human body. It is important to emphasize that an X-ray scanner allows the generation of different images, by considering different angles around the body part which is undergo analysis. Once processed by the dedicated computer, the output is a collection of the cross-sectional images, often referred as slices.

3-D MR images are an important source of information, in different medical applications and, especially, in medical

diagnosis (ranging from neuroimaging to oncology). MR images are often preferred. In fact, in the case in which, both CT and MR images, produce the same information, these latter techniques are preferred, since MR acquisitions do not use any ionizing radiation. From the other hand, in presence of subjects with cardiac pacemakers and/or metallic foreign bodies, MR techniques cannot be applied.

Medical data need to be managed in an efficient and effective manner and data compression techniques are essential in order to solve the transmission and storage problems.

For medical images, lossless compression is often required and, in many situations, indispensable. In fact, these data are precious or often obtained by means of unrepeatable medical exams.

Lossy compression techniques could sometime be considered, but it is necessary take into account that that the information lost might lead to incorrect diagnosis or it could affect the reanalysis of data.

In this paper, we consider lossless predictive techniques. We have focused on multidimensional medical image sequences (3-D computed tomography images, functional resonance magnetic images), which have considerable space memory requirements (many hundreds of megabytes/gigabytes per acquisition).

This paper introduces a multidimensional, configurable, predictive structure that can be used for the compression of multidimensional medical images.

The predictor we propose is scalable, adjustable, and adaptive and we present experimental evidences of its performance on multidimensional medical images: 3-D Computed Tomography (CT) and 3-D Magnetic Resonance (MR).

This paper is organized as follows: Section 2 describes the predictive structure. Section 3 reports our experimental results. Section 4 highlights our conclusions and outlines future research directions.

## II. PREDICTIVE CODING FOR MULTIDIMENSIONAL IMAGES

The predictive model we propose is based on the least squares optimization technique. In order to perform the prediction of the current sample, a prediction context, composed by the neighboring samples of the current component and one (or more) reference component(s), is used. The reference component(s) can be of different dimension(s), with respect to the current component. Therefore the

R. Pizzolante and B. Carpentieri are with the Dipartimento di Informatica, Università di Salerno, I-84084 Fisciano (SA) – Italia. (phone: +39 089969500; fax: +39 089969600/1; e-mail: [rpizzolante@unisa.it](mailto:rpizzolante@unisa.it), [bc@dia.unisa.it](mailto:bc@dia.unisa.it)).

prediction is achieved by using a multidimensional prediction context.

Without loss of generality, for the following definitions, we assume that the multidimensional image which we have to compress, has the following size:  $\langle M_1, M_2, \dots, M_{N-2}, X, Y \rangle$ , where  $X$  and  $Y$  are respectively the width and the height of the bi-dimensional components and  $M_f$  is the size of the  $f$ -th dimension ( $1 \leq f \leq N - 2$ ). A specific bi-dimensional component can be univocally identified through a vector of  $N - 2$  elements:  $[p_1, p_2, \dots, p_{N-2}]$ , where  $p_i \in \{1, 2, \dots, M_i\}$ .

For example, if we consider a three-dimensional image,  $\langle Z, X, Y \rangle$  the dataset is composed of  $Z$  components (among the third dimension), where each component has respectively width  $X$  and height  $Y$ .

As we outlined above, our predictive model uses one or more reference components, which will be specified through the Sets of References or (References Set).

If the current sample has coordinates  $(m_1, m_2, \dots, m_{N-2}, x, y)$  (where  $1 \leq x \leq X$  and  $1 \leq y \leq Y$ ), for each of the  $N - 2$  dimension, we define a *references set*, denoted as:

$$R_i = \{r_1^i, r_2^i, \dots, r_{t_i}^i\}, \text{ for } i \in \{1, 2, \dots, N - 2\}, \text{ where } r_j^i \in \{1, 2, \dots, M_i\} \cup \{-1, -2, \dots, -M_i\}, t_i = |R_i|, 1 \leq j \leq t_i,$$

and  $\left| \bigcup_{i=1}^{N-2} R_i \right| > 0$ .

Such references sets are univocally set up at the beginning of the algorithm and are used in the prediction step.

A generic element  $r_j^i \in R_i$  ( $1 \leq i \leq N - 2$ ), denotes a specific bi-dimensional component. We will use the following notation: if  $r_j^i > 0$ , then the denoted component is the one identified through the vector  $[m_1, m_2, \dots, m_{i-1}, r_j^i, m_{i+1}, \dots, m_{N-2}]$ , or, if  $r_j^i < 0$ , then the denoted component is the one identified through the vector  $[m_1, m_2, \dots, m_{i-1}, m_i - |r_j^i|, m_{i+1}, \dots, m_{N-2}]$ .

The proposed predictive model is based on the least squares optimization technique. The prediction is formed by using the current component and all the (valid) components of the references sets.

In order to refer to a sample without the use of its coordinates, we define an *enumeration*. Its main objective is the relative indexing among all the samples (or a subset of them) of the same component. In particular, by fixing a sample, namely the reference sample, all the other samples of the component will be indexed with respect to it. Therefore, in this manner, it is possible to address a sample by using its relative index. The relative indexing of the samples is used for the definition of the multidimensional prediction context involved by our predictive model.

Let  $E$  denotes a 2-D enumeration, which has as objective the relative indexing of the samples in a bi-dimensional context, with respect to a specific reference sample. The fundamental requisites that the enumeration  $E$  needs to satisfy

are that the specified reference sample has 0 as index and that any two samples (with different coordinates) do not have the same index.

Let  $x_j^{(e)}(r_s^j)$  (where  $r_s^j \in R_j$ ) denotes the  $e$ -th sample in the bi-dimensional context according to the enumeration  $E$  with respect to the sample with coordinate  $(m_1, m_2, \dots, m_{j-1}, r_s^j, m_{j+1}, \dots, m_{N-2}, x, y)$  when  $r_s^j > 0$ , or  $(m_1, m_2, \dots, m_{j-1}, m_j - |r_s^j|, m_{j+1}, \dots, m_{N-2}, x, y)$  when  $r_s^j < 0$ .

Finally, let  $x^{(e)}$  denotes the  $e$ -th sample, according to the enumeration  $E$ , with respect to the current sample. Notice that  $x^{(0)}$  denotes precisely the current sample.

The  $T$ -order prediction (where  $T = \sum_{i=1}^{N-2} t_i = \sum_{i=1}^{N-2} |R_i|$ ) of the current sample  $x^{(0)}$  is obtained by:

$$\hat{x}^{(0)} = \sum_{i=1}^{N-2} \sum_{j=1}^{t_i} \alpha_i^j \cdot x_i^{(0)}(r_j^i). \quad (1)$$

The  $\alpha_0 = [\alpha_1^1, \dots, \alpha_{t_1}^1, \dots, \alpha_1^2, \dots, \alpha_{t_2}^2, \dots, \alpha_1^{N-2}, \dots, \alpha_{t_{N-2}}^{N-2}]^t$  coefficients are chosen to minimize the energy of the prediction error:

$$P = \sum_{i=1}^H (x^{(i)} - \hat{x}^{(i)})^2 \quad (2)$$

$H$  indicates the number of samples used, for the current and for each of the components specified in the references sets. Thus,  $H \cdot (T + 1) + T$  samples are used for the prediction.

The coefficients  $\alpha_0$  are obtained by using the optimal linear prediction method, as in [25].

We can rewrite the equation (2) in the form:

$$P = (C\alpha - X)^t \cdot (C\alpha - X),$$

by using matrix notation.

The linear system is obtained, as in [25], by taking the derivate of the equation (2) with respect to  $\alpha$ , and by setting it to zero.

$$(C^t C) \alpha_0 = (C^t X). \quad (3)$$

Thus, by computing the coefficients  $\alpha_0$ , which solve the linear system (3), it is possible to determinate the prediction of the current sample,  $\hat{x}^{(0)}$ , by using equation (1).

The prediction error

$$e = [x^{(0)} - \hat{x}^{(0)}] \quad (4)$$

can then be sent to an entropy encoder.

It is important to outline that  $Hx(T+1) + H$  samples are used to achieve the prediction. Our predictive structure involves only by past information: there is no need to send any side information to the decompression algorithm.

The computational complexity of the prediction is related to the two configurable parameters:  $H$  and the Sets of

References. It is possible to model the multidimensional prediction context by specifying its wideness and the number of the reference components. By doing this it is possible either to define a prediction context which can minimize the use of the computational resources or to refine the accurateness of the prediction by using more computational resources.

In some situations, our predictive structure can be ineffective. In particular, when the linear system of equations (3) cannot be solved because it has no solutions or infinitely many solutions. In such scenarios, which we referred as *exceptions*, the predictive structure is not able to perform the prediction.

In presence of a sample that cannot be predicted through the proposed predictive structure (because an exception is verified), an alternative predictive structure (as for instance Median Predictor, etc.) shall be used.

### III. EXPERIMENTAL RESULTS

We have tested our prediction model by implementing a predictive-based compression scheme, and then we have experimented this algorithm on two different types of multidimensional medical images: 3-D computed tomography images and 3-D magnetic resonance images.

The algorithm predicts the current sample by using the previously coded samples. In this way, it is possible to have a consistent prediction for both the compression and the decompression algorithm.

After the prediction step, the prediction error is obtained by the encoder as a difference between the current sample and its prediction.

Finally, the prediction error can be encoded by using an entropy or a statistical coder. In our tests, we have used as error encoder: PAQ8 [10], Prediction by Partial Matching with Information Inheritance (PPMd or PPMII) [26].

The algorithm uses the 2-D Linearized Median Predictor (2D-LMP) [21], for all the components which have no component references, and our multidimensional predictive structure, for all the other components.

In order to define the prediction context, we need to enumerate the neighboring pixels of  $X$  in the current and in the previous bands.

For these reasons, we define an enumeration that depends on a distance  $d$ , defined as:

$$d((z, u, v), (z, w, z)) = \sqrt{(u - w)^2 + (v - z)^2}$$

When more pixels have the same indices, it is possible to reassign the indices of these pixels in clockwise order with respect to  $X$ .

To improve the readability, we used the mnemonic name of the dimension instead of its index for the references sets. For example,  $R_Z$  indicates the reference set for the  $Z$  dimension.

#### 3.1. 3-D Computed Tomography Images

We have performed experiments on a the test set described in [21], composed by four 3-D CT images, in which each sample is stored by using 8 bits. For the coding of prediction errors, which we have mapped similarly to [15], then we have used PAQ8 and we have managed the exceptions with the 3-D Differences-based Linearized Median Predictor (3D-DLMP) [21].

The following tables summarize the results we have obtained on the four CT images in terms of bits-per-sample (BPS) by using different configurations for the  $H$  parameter and the references set. The results are compared with other state-of-the-art techniques.

Our approach outperforms, in average, all the other state-of-the-art techniques, as it is possible to see from Table 5.

| Methods / Images                    |                            | CT_skull        |
|-------------------------------------|----------------------------|-----------------|
| Dimensions                          |                            | <192, 256, 256> |
| <i>Proposed</i><br>\\<br>Parameters | $H=32, R_Z=\{-1, -2, -3\}$ | 1.4836          |
|                                     | $H=16, R_Z=\{-1, -2, -3\}$ | 1.5309          |
|                                     | $H=8, R_Z=\{-1, -2, -3\}$  | 1.6258          |
|                                     | $H=32, R_Z=\{-1, -2\}$     | 1.5393          |
|                                     | $H=16, R_Z=\{-1, -2\}$     | 1.5688          |
|                                     | $H=8, R_Z=\{-1, -2\}$      | 1.6196          |
| 3D-ESCOT [28]                       |                            | 1.8350          |
| MILC [21]                           |                            | 2.0306          |
| AT-SPIHT [6]                        |                            | 1.9180          |
| 3D-CB-EZW [3]                       |                            | 2.0095          |
| DPCM+PPMd [1]                       |                            | 2.1190          |
| 3D-SPIHT [28]                       |                            | 1.9750          |
| 3D-EZW [3]                          |                            | 2.2251          |
| JPEG-LS [4]                         |                            | 2.8460          |

**Table 1:** Experimental results obtained on CT skull

| Methods / Images                    |                            | CT_wrist        |
|-------------------------------------|----------------------------|-----------------|
| Dimensions                          |                            | <176, 256, 256> |
| <i>Proposed</i><br>\\<br>Parameters | $H=32, R_Z=\{-1, -2, -3\}$ | 0.8979          |
|                                     | $H=16, R_Z=\{-1, -2, -3\}$ | 0.9290          |
|                                     | $H=8, R_Z=\{-1, -2, -3\}$  | 1.0042          |
|                                     | $H=32, R_Z=\{-1, -2\}$     | 0.9527          |
|                                     | $H=16, R_Z=\{-1, -2\}$     | 0.9737          |
|                                     | $H=8, R_Z=\{-1, -2\}$      | 1.0110          |
| 3D-ESCOT [28]                       |                            | 1.0570          |
| MILC [21]                           |                            | 1.0666          |
| AT-SPIHT [6]                        |                            | 1.1150          |
| 3D-CB-EZW [3]                       |                            | 1.1393          |
| DPCM+PPMd [1]                       |                            | 1.0290          |
| 3D-SPIHT [28]                       |                            | 1.1720          |
| 3D-EZW [3]                          |                            | 1.2828          |
| JPEG-LS [4]                         |                            | 1.6531          |

**Table 2:** Experimental results obtained on CT wrist

3.2 3-D Magnetic Resonance Images

We have performed similar experiments also for the four MR images commonly used for testing in the literature.

As for the CT images the following tables show that our approach outperform the current state of the art algorithms.

| Methods / Images<br>Dimensions |                            | CT_carotid<br><64, 256, 256> |
|--------------------------------|----------------------------|------------------------------|
| Proposed<br>Parameters         | $H=32, R_z=\{-1, -2, -3\}$ | 1.2783                       |
|                                | $H=16, R_z=\{-1, -2, -3\}$ | 1.2976                       |
|                                | $H=8, R_z=\{-1, -2, -3\}$  | 1.3421                       |
|                                | $H=32, R_z=\{-1, -2\}$     | 1.3363                       |
|                                | $H=16, R_z=\{-1, -2\}$     | 1.3448                       |
|                                | $H=8, R_z=\{-1, -2\}$      | 1.3496                       |
|                                | 3D-ESCOT [28]              | 1.3470                       |
|                                | MILC [21]                  | 1.3584                       |
|                                | AT-SPIHT [6]               | 1.4790                       |
| 3D-CB-EZW [3]                  | 1.3930                     |                              |
| DPCM+PPMd [1]                  | 1.4710                     |                              |
| 3D-SPIHT [28]                  | 1.4340                     |                              |
| 3D-EZW [3]                     | 1.5069                     |                              |
| JPEG-LS [4]                    | 1.7388                     |                              |

Table 3: Experimental results obtained on CT carotid

| Methods / Images<br>Dimensions |                            | CT_aperts<br><96, 256, 256> |
|--------------------------------|----------------------------|-----------------------------|
| Proposed<br>Parameters         | $H=32, R_z=\{-1, -2, -3\}$ | 0.7283                      |
|                                | $H=16, R_z=\{-1, -2, -3\}$ | 0.7350                      |
|                                | $H=8, R_z=\{-1, -2, -3\}$  | 0.7587                      |
|                                | $H=32, R_z=\{-1, -2\}$     | 0.7265                      |
|                                | $H=16, R_z=\{-1, -2\}$     | 0.7271                      |
|                                | $H=8, R_z=\{-1, -2\}$      | 0.7349                      |
|                                | 3D-ESCOT [28]              | 0.8580                      |
|                                | MILC [21]                  | 0.8190                      |
|                                | AT-SPIHT [6]               | 0.9090                      |
| 3D-CB-EZW [3]                  | 0.8923                     |                             |
| DPCM+PPMd [1]                  | 0.8670                     |                             |
| 3D-SPIHT [28]                  | 0.9980                     |                             |
| 3D-EZW [3]                     | 1.0024                     |                             |
| JPEG-LS [4]                    | 1.0637                     |                             |

Table 4: Experimental results obtained on CT aperts

| Methods / Images<br>Dimensions |                            | Average |
|--------------------------------|----------------------------|---------|
| Proposed<br>Parameters         | $H=32, R_z=\{-1, -2, -3\}$ | 1.0970  |
|                                | $H=16, R_z=\{-1, -2, -3\}$ | 1.1231  |
|                                | $H=8, R_z=\{-1, -2, -3\}$  | 1.1827  |
|                                | $H=32, R_z=\{-1, -2\}$     | 1.1387  |
|                                | $H=16, R_z=\{-1, -2\}$     | 1.1536  |
|                                | $H=8, R_z=\{-1, -2\}$      | 1.1788  |
|                                | 3D-ESCOT [28]              | 1.2743  |
|                                | MILC [21]                  | 1.3187  |
|                                | AT-SPIHT [6]               | 1.3553  |
| 3D-CB-EZW [3]                  | 1.3585                     |         |
| DPCM+PPMd [1]                  | 1.3715                     |         |
| 3D-SPIHT [28]                  | 1.3948                     |         |
| 3D-EZW [3]                     | 1.5043                     |         |
| JPEG-LS [4]                    | 1.8254                     |         |

Table 5: Average experimental results obtained on the four CT images.

| Methods / Images<br>Dimensions |                            | MR_liver_t1<br><48, 256, 256> |
|--------------------------------|----------------------------|-------------------------------|
| Proposed<br>Parameters         | $H=32, R_z=\{-1, -2, -3\}$ | 1.8511                        |
|                                | $H=16, R_z=\{-1, -2, -3\}$ | 1.8850                        |
|                                | $H=8, R_z=\{-1, -2, -3\}$  | 1.9894                        |
|                                | $H=32, R_z=\{-1, -2\}$     | 1.8996                        |
|                                | $H=16, R_z=\{-1, -2\}$     | 1.9089                        |
|                                | $H=8, R_z=\{-1, -2\}$      | 1.9471                        |
| 3D-ESCOT                       | 2.0760                     |                               |
| MILC                           | 2.1968                     |                               |
| 3D-SPIHT                       | 2.2480                     |                               |
| 3D-CB-EZW                      | 2.2076                     |                               |
| DPCM+PPMd                      | 2.3900                     |                               |
| 3D-EZW                         | 2.3743                     |                               |
| JPEG-LS                        | 3.1582                     |                               |

Table 6: Experimental results obtained on MR liver\_t1

| Methods / Images<br>Dimensions |                            | MR_liver_t2e1<br><48, 256, 256> |
|--------------------------------|----------------------------|---------------------------------|
| Proposed<br>Parameters         | $H=32, R_z=\{-1, -2, -3\}$ | 1.2539                          |
|                                | $H=16, R_z=\{-1, -2, -3\}$ | 1.2783                          |
|                                | $H=8, R_z=\{-1, -2, -3\}$  | 1.3360                          |
|                                | $H=32, R_z=\{-1, -2\}$     | 1.3101                          |
|                                | $H=16, R_z=\{-1, -2\}$     | 1.3232                          |
|                                | $H=8, R_z=\{-1, -2\}$      | 1.3482                          |
| 3D-ESCOT                       | 1.5100                     |                                 |
| MILC                           | 1.7590                     |                                 |
| 3D-SPIHT                       | 1.6700                     |                                 |
| 3D-CB-EZW                      | 1.6591                     |                                 |
| DPCM+PPMd                      | 2.0250                     |                                 |
| 3D-EZW                         | 1.8085                     |                                 |
| JPEG-LS                        | 2.3692                     |                                 |

Table 7: Experimental results obtained on MR liver\_t2e1

| Methods / Images<br>Dimensions |                            | MR_sag_head<br><48, 256, 256> |
|--------------------------------|----------------------------|-------------------------------|
| Proposed<br>Parameters         | $H=32, R_z=\{-1, -2, -3\}$ | 1.4890                        |
|                                | $H=16, R_z=\{-1, -2, -3\}$ | 1.5311                        |
|                                | $H=8, R_z=\{-1, -2, -3\}$  | 1.6020                        |
|                                | $H=32, R_z=\{-1, -2\}$     | 1.5477                        |
|                                | $H=16, R_z=\{-1, -2\}$     | 1.5737                        |
|                                | $H=8, R_z=\{-1, -2\}$      | 1.6094                        |
| 3D-ESCOT                       | 1.9370                     |                               |
| MILC                           | 2.0975                     |                               |
| 3D-SPIHT                       | 2.0710                     |                               |
| 3D-CB-EZW                      | 2.2846                     |                               |
| DPCM+PPMd                      | 2.1270                     |                               |
| 3D-EZW                         | 2.3883                     |                               |
| JPEG-LS                        | 2.5567                     |                               |

Table 8: Experimental results obtained on MR sag\_head

| Methods / Images<br>Dimensions |                            | MR_ped_chest<br><64, 256, 256> |
|--------------------------------|----------------------------|--------------------------------|
| Proposed<br>//<br>Parameters   | $H=32, R_z=\{-1, -2, -3\}$ | 1.2920                         |
|                                | $H=16, R_z=\{-1, -2, -3\}$ | 1.3498                         |
|                                | $H=8, R_z=\{-1, -2, -3\}$  | 1.4669                         |
|                                | $H=32, R_z=\{-1, -2\}$     | 1.3740                         |
|                                | $H=16, R_z=\{-1, -2\}$     | 1.4053                         |
|                                | $H=8, R_z=\{-1, -2\}$      | 1.4694                         |
|                                | 3D-ESCOT                   | 1.6180                         |
|                                | MILC                       | 1.6556                         |
|                                | 3D-SPIHT                   | 1.7420                         |
| 3D-CB-EZW                      | 1.8705                     |                                |
| DPCM+PPMd                      | 1.6890                     |                                |
| 3D-EZW                         | 2.0499                     |                                |
| JPEG-LS                        | 2.9282                     |                                |

Table 9: Experimental results obtained on MR ped\_chest

| Methods / Images<br>Dimensions |                            | Average |
|--------------------------------|----------------------------|---------|
| Proposed<br>//<br>Parameters   | $H=32, R_z=\{-1, -2, -3\}$ | 1.4715  |
|                                | $H=16, R_z=\{-1, -2, -3\}$ | 1.5111  |
|                                | $H=8, R_z=\{-1, -2, -3\}$  | 1.5986  |
|                                | $H=32, R_z=\{-1, -2\}$     | 1.5329  |
|                                | $H=16, R_z=\{-1, -2\}$     | 1.5528  |
|                                | $H=8, R_z=\{-1, -2\}$      | 1.5935  |
|                                | 3D-ESCOT                   | 1.7853  |
|                                | MILC                       | 1.9272  |
|                                | 3D-SPIHT                   | 1.9328  |
| 3D-CB-EZW                      | 2.0055                     |         |
| DPCM+PPMd                      | 2.0578                     |         |
| 3D-EZW                         | 2.1553                     |         |
| JPEG-LS                        | 2.7531                     |         |

Table 10: Average experimental results obtained on the four MR images.

#### IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a Multidimensional Predictive Model that can be used for lossless compression of multidimensional medical image. We have experimentally tested our method on 3-D magnetic resonance (MR) and 3-D computed tomography (CT) images.

Future work will include the testing of our approach on 4-D and 5-D functional Magnetic Resonance Imaging (fMRI) data, the usage of the model in a lossy codec, and deeper experimentation on lossless compression by using other N-D data (eg. 4-D ultrasound images, etc.).

#### REFERENCES

[1] Ait-Aoudia, S.; Benhamida, F.; and Yousfi, M., "Lossless Compression of Volumetric Medical Data", Lecture Notes in Computer Science, vol. 4263/2006, pp. 563-571, 2006.

[2] Aron AR, Behrens TE, Smith S, Frank MJ, Poldrack RA., "Triangulating a Cognitive Control Network Using Diffusion-Weighted Magnetic Resonance Imaging (MRI) and Functional MRI", The Journal of Neuroscience, 4 April 2007, 27(14): 3743-3752

[3] Bilgin A., Zweig G., and Marcellin M.W., "Three-Dimensional Image Compression with Integer Wavelet", Applied Optics, vol. 39, no. 11, pp. 1799-1814, April, 2000.

[4] Carpentieri B., Weinberger M., Seroussi G., "Lossless Compression of Continuous Tone Images", Proceeding of IEEE, vol. 88, no. 11, pp. 1797-1809, November, 2000.

[5] Carpentieri, B., Storer, J.A., Motta, G., Rizzo F., "Compression of Hyperspectral Imagery", Proceedings of IEEE Data Compression Conference (DCC '03), Snowbird, UT, USA, pp. 317-324, 25-27 March 2003.

[6] Cho S., Kim D. and Pearlman W.A., "Lossless Compression of Volumetric Medical Images with Improved Three-Dimensional SPIHT Algorithm", Journal of Digital Imaging, vol. 17, no. 1, pp. 57-63, March, 2004.

[7] fMRI Wikipedia English Page, Available on: <http://en.wikipedia.org/wiki/Fmri> (Accessed on Oct. 2013).

[8] Galoppo, N.; Govindaraju, N.K.; Henson, M.; Manocha, D., "LU-GPU: Efficient Algorithms for Solving Dense Linear Systems on Graphics Hardware," Supercomputing, 2005. Proceedings of the ACM/IEEE SC 2005 Conference, vol., no., pp.3,3, 12-18 Nov. 2005

[9] Golub G.H., Van Loan C.F., "Matrix Computations, 3rd ed. Baltimore", MD: The Johns Hopkins Univ. Press, 1996.

[10] Knoll, B.; de Freitas, N., "A Machine Learning Perspective on Predictive Coding with PAQ8," Data Compression Conference (DCC), 2012, vol., no., pp.377,386, 10-12 April 2012.

[11] Lalgudi, H.G.; Bilgin, A.; Marcellin, M.W.; Nadar, M.S., "Compression of Multidimensional Images Using JPEG2000," Signal Processing Letters, IEEE, vol.15, no., pp.393,396, 2008. R. W. Lucky, "Automatic equalization for digital communication," Bell Syst. Tech. J., vol. 44, no. 4, pp. 547-588, Apr. 1965.

[12] Magli E., Olmo G., Quacchio E., "Optimized onboard lossless and near-lossless compression of hyperspectral data using CALIC", Geoscience and Remote Sensing Letters, IEEE, vol.1, no.1, pp.21,25, Jan. 2004.

[13] Martinez-Alonso, R.; Mino, K.; Torres-Lucio, D., "Array Processors Designed with VHDL for Solution of Linear Equation Systems Implemented in a FPGA," Electronics, Robotics and Automotive Mechanics Conference (CERMA), 2010, vol., no., pp.731,736, Sept. 28 2010-Oct. 1 2010.

[14] Mielikainen J., "Lossless compression of hyperspectral images using lookup tables", IEEE Signal Process. Letters, vol. 13, no. 3, pp. 157-160, Mar. 2006.

[15] Motta G., Storer J. A., and Carpentieri B., "Lossless Image Coding via Adaptive Linear Prediction and Classification", Proceedings of the IEEE, vol. 88, no. 11, pp. 1790-1796, November, 2000.

[16] Motta, G.; Rizzo, F.; Storer, J.A., "Hyperspectral Data Compression", Springer Science: Berlin, Germany, 2006.

[17] Muñoz-Gómez, J.; Bartrina-Rapesta, J.; Blanes, I.; Jiménez-Rodríguez, L.; Aulí-Llinàs, F.; Serra-Sagrístà, J.; "4D remote sensing image coding with JPEG2000", Proc. SPIE 7810, SDCCP VI, 78100X, August 24, 2010.

[18] NASA AVIRIS Page: Available on: <http://aviris.jpl.nasa.gov/> (Accessed on Oct. 2013).

[19] NODC Overview, Available on: <http://www.nodc.noaa.gov/about/overview.html> (Accessed on Oct. 2013).

[20] OpenfMRI Site, Available on: <https://openfmri.org> (Accessed on Oct. 2013).

[21] Pizzolante, R.; Carpentieri, B., "Lossless, low-complexity, compression of three-dimensional volumetric medical images via linear prediction", Digital Signal Processing (DSP), 2013, pp.1,6, 1-3 July 2013.

[22] Pizzolante, R.; Carpentieri, B., "Visualization, Band Ordering and Compression of Hyperspectral Images", Algorithms 2012, 5, 76-97.

[23] Rissanen, J., "Generalized Kraft inequality and arithmetic coding", IBM J. Res. Develop., vol. 20 (3), pp. 198-203, May 1976.

[24] Rissanen, J., Langdon, Jr. G. G., "Universal modeling and coding", IEEE Trans. Inform. Theory, vol. IT-27, pp. 12-23, Jan. 1981.

[25] Rizzo F., Carpentieri B., Motta G., Storer J.A., "Low-complexity lossless compression of hyperspectral imagery via linear prediction", Signal Processing Letters, IEEE, vol.12, no.2, pp. 138-141, Feb. 2005.

[26] Shkarin D., "PPM: one step to practicality", Data Compression Conference (DCC) 2002, pp. 202-211, April, 2002.

[27] Taubman, D.; Marcellin, M.W., "JPEG2000 Image Compression Fundamentals, Standards and Practice", The Springer International Series in Engineering and Computer Science, Vol. 642, 2002.

[28] Xiong, Z.; Wu, X.; Cheng, S.; Jianping H., "Lossy-to-lossless compression of medical volumetric data using three-dimensional integer wavelet transforms," IEEE Trans. on Medical Imaging, vol.22, no.3, pp.459,470, 2003.

# Does Time Pressure Induce Tunnel Vision? An examination with the Eriksen Flanker Task by applying the Hierarchical Drift Diffusion Model

Nico Assink, Rob H. J. van der Lubbe, and Jean-Paul Fox

*Abstract*— Mental stress is often thought to induce a phenomenon denoted as tunnel vision, which may be characterized as a shrinkage in the size of the attentional focus. This seems to imply that potentially relevant information is not taken into account while making a certain decision. In experimental settings, an effective way to induce mental stress is the use of time pressure by employing strict response deadlines. We decided to use the Eriksen flanker task to examine whether time pressure induces tunnel vision. The effect of peripheral flanker stimuli on both response speed and accuracy was compared between low and high time pressure conditions in three experiments. Instead of focusing solely on the speed and accuracy of responses, we decided to use the hierarchical drift diffusion model to determine the values of relevant parameters that describe the underlying decision process: the response criterion ( $\alpha$ ) and the drift rate ( $v$ ). The results consistently revealed that time pressure reduced the response criterion. Importantly, incongruent flankers reduced the drift rate under high time pressure as compared to low time pressure. The latter pattern of results is not in line with the idea that mental stress induces tunnel vision.

*Keywords*—: drift rate, Eriksen flanker task, hierarchical drift diffusion model, response criterion, time pressure, tunnel vision,

## I. INTRODUCTION

Imagine you are driving your car in a city when suddenly the car in front of you hits the brakes. You have to react to this unexpected event and better do it quickly. Your heart rate increases, you clench the steering wheel and your eyes widen while your foot releases the gas pedal and hits the brake. The only thing you see is the car in front of you and its brake lights. Thanks to your physical reactions to the sudden threat and your focused attention a crash is averted and the stream of cars starts to pick up speed again. Then, totally unexpected, a car crashes into the right side of your car. As you calm down

Nico Assink is with the Department of Cognitive Psychology and Ergonomics, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. (e-mail: nico.assink@gmail.com).

Rob H. J. van der Lubbe is with the Department of Cognitive Psychology and Ergonomics, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. (e-mail: r.h.j.vanderlubbe@utwente.nl).

Jean-Paul Fox is with the Department of Research Methodology, Measurement, and Data Analysis, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. (e-mail: j.p.fox@utwente.nl).

and realize what just happened, you wonder how you ever could have missed that red traffic light.

In a stressful situation such as described above a phenomenon called tunnel vision seems to occur. Information from the attended part of the visual field is still fully processed, but visual clues from other parts of the visual field that would otherwise be detected remain completely unnoticed. Thus, tunnel vision may be characterized as a shrinkage in the size of the attentional focus. This supposed change in visual attention as a result of mental stress is often taken as a fact in applied settings, but the evidence from research is not that conclusive. The aim of the current research is to answer the question whether stress manipulated by varying time pressure induces tunnel vision.

Behavioral studies reported some support for tunnel vision as a result of different stressors. Reference [1] shows observed reduced performance on a secondary, peripheral signal detection task in hot and humid conditions. However, the results of an experiment reported by [2], who used an evaluative observer to induce stress, only partially confirmed the view that stress induces tunnel vision. In Dirkin's experiment, [2], participants had to identify the number of illuminated lights on any of three display panels, with one centrally located panel and two peripherally located panels placed at an angle of 70° to the left and right of the subject's median. The identification of the lights on the peripheral panels constituted the primary task, and identification of the lights on the central panel was the secondary task. Under stress, the performance on the primary task improved, however, the hypothesized decrease in performance on the secondary task was not found.

Results from other electrophysiological studies, in which time pressure was used as a stressor, do not match well with the idea that tunnel vision occurs as a result of stress. Reference [3] shows event-related potentials (ERPs) derived from the electroencephalogram (EEG) to examine the mechanisms underlying speed-accuracy trade-off (SAT). Participants in their study performed a choice reaction time (RT) task known as the Eriksen flanker task (e.g., [4]). In this task, participants have to respond to the identity of a centrally presented target stimulus with a left or right button press as fast and as accurately as possible. The target is accompanied by irrelevant flanker stimuli. On congruent trials, the flankers signal the same response as the target while on incongruent trials the flankers correspond with the opposite response. Participants typically respond faster and more accurate on congruent than on incongruent trials, indicating an inability to completely

ignore the flankers. In the study of Osman et al. task instructions varied between blocks, emphasizing either speed or accuracy. Instructions emphasizing speed resulted in a later onset of the response-locked lateralized readiness potential (r-LRP), but did not affect the onset latency of the stimulus-locked version of this potential (s-LRP). These results indicate that speed-accuracy instructions only affected the portion of RT following the start of motor preparation. Osman et al. also determined effects on the P300 ERP component, which is thought to be primarily affected by changes in early processing stages such as stimulus evaluation. The peak latency of the P300 potential was affected by target-flanker congruency, with an earlier peak on congruent than on incongruent trials. The speed-accuracy instructions, however, did not affect the latency of this peak, adding support to the conclusion that only late processes are affected by speed-accuracy instructions.

Reference [5] the influence of time pressure in a simple response task is examined, a choice-by-location task and the Simon task by varying response deadlines. In both the choice-by-location task and the Simon task, they observed that time pressure had no influence on the s-LRP while it affected the r-LRP, which corresponds with the results of [3] with the Eriksen task. Another lateralized EEG potential, the posterior contralateral negativity (PCN/N2pc) was used to provide more information about the influence of time pressure on earlier pre-motoric processes. The onset of the PCN may be used as an index for the start of discriminative processing of the relevant aspect of the stimuli. A change in onset or peak latency of this potential caused by different levels of time pressure would indicate that attentional orienting was affected. No such effects were observed. Together, these findings accord with the view that time pressure does not affect early attentional processes, but only later motor processes. However, the Eriksen flanker task seems more appropriate for demonstrating the presence or absence of tunnel vision, as successful execution of this task seems to depend on the size of the attentional window.

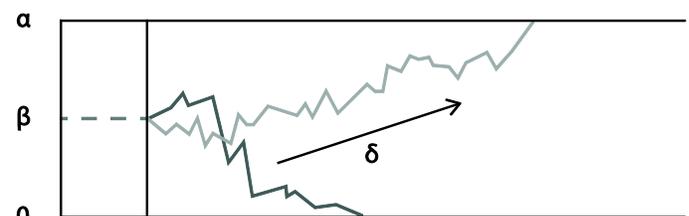
In conclusion, there appears to be a discrepancy between the results of the EEG studies of [3] and [5], and the long held assumption based on earlier behavioral studies that stress induces tunnel vision. In an attempt to bridge the gap between studies using behavioral measures and studies using electrophysiological measures, we measured overt behavioral measures (speed and accuracy), and related them to properties of a model that has been proposed to reflect the underlying neurophysiological processes. More specifically, we incorporated both speed and accuracy information to determine the properties of the underlying response selection process by using the Hierarchical Drift Diffusion Model (HDDM, see [6]).

In the current study, an arrowhead version of the Eriksen flanker task was employed. The high overlap between stimuli and responses is known to result in a strong response conflict (e.g., see [7]). In this task version, participants are instructed to respond as fast and accurately as possible by pushing a left button if the centrally presented target arrow points to the left and a right button if the target points to the right. The target is flanked by two distractors on both sides that can either point in the same (congruent) or opposite direction (incongruent) as

the target. In a neutral condition, two parallel horizontal lines were used as flanker stimuli. Incongruent flankers reduce the speed and accuracy of the responses, despite clear instructions informing the participant to attend only to the identity of the central target stimuli. Reference [4] interpreted this as an inability to completely ignore the irrelevant flankers. This congruency effect makes the task well suited to test for the presence of tunnel vision. Namely, in the case of tunnel vision the processing of the flanker stimuli should diminish leading to a reduction of the congruency effect.

The congruency effect can be manifested in both speed and accuracy. This poses a challenge in interpreting RT and proportion correct (PC) because of SAT. Participants may respond faster in a certain condition at the expense of accuracy or vice versa ([8],[9]). Thus, if speed and accuracy change in opposite directions, it may be difficult to interpret specific performance differences between conditions. If a manipulation results in a great increase in speed, but simultaneously in more errors, the question may be raised whether the manipulation made the task more or less difficult, as speed and accuracy cannot be directly translated into each other. The HDDM overcomes this problem as it allows for the comparison of different properties of the underlying decision process.

The HDDM is a prominent sequential sampling model for two-choice decisions. It assumes that evidence for a specific response accumulates over time from a noisy input signal ([6],[10],[11]). When enough evidence for a specific response has been accumulated the response will be executed. Figure 1 shows a graphical representation of the process and its parameters.



**Fig. 1** A graphical representation of the hierarchical drift diffusion model (HDDM).  $\alpha$  = response criterion;  $\delta$  = drift rate per individual trial;  $\beta$  = bias. The history of two possible decision processes is shown, one reaching the top boundary, leading to a correct response, and one reaching the bottom boundary, leading to an incorrect response.

Evidence is thought to accumulate with an average rate per trial called the drift rate ( $v$ ). The amount of evidence needed for a response is indicated by the boundary separation or response criterion ( $\alpha$ ). The initial starting point of the process is determined by  $\beta$  as a proportion of  $\alpha$ . This value amounts to 0.5 in the case of no bias. The top boundary represents the evidence required to give a correct response while the bottom boundary indicates the evidence that will lead to an incorrect response. Another parameter  $\tau$  represents the non-decision time, which is the time needed for all processes apart from the decision process, such as sensory processing and physically executing the response.

Our main interest is in effects on the parameters for the response criterion ( $\alpha$ ), and the drift rate ( $v$ ). A high response

criterion will result in slower but more accurate responses. This means that a change in  $\alpha$  might explain the SAT phenomenon (e.g., see [12],[13]). For example, in a high time pressure condition, the size of the boundary separation may be reduced, enabling the participant to respond more quickly, but at the expense of accuracy. The value of the drift rate parameter ( $\nu$ ) represents the average rate of evidence accumulation on a trial. Here, evidence means the amount of information regarding a specific response. On congruent trials all stimuli are likely to contribute evidence towards the correct response. On neutral trials, evidence can be sampled from the target, while the flanker stimuli only provide noise. On incongruent trials, sampling information from the flanker stimuli may actually reduce the amount of accumulated evidence. Thus, we may expect the drift rate to be largest on congruent trials, intermediate on neutral trials, and smallest on incongruent trials. If flanker stimuli would be completely ignored, which might occur in an extreme version of tunnel vision, then evidence should accumulate at the same rate in each condition, as the target always conveys the same amount of information about the correct response. If flankers are partially ignored due to tunnel vision, this should decrease the rate of evidence accumulation on congruent trials, while it should increase the rate on incongruent trials. The difference in drift rate between congruent and incongruent trials thus represents a congruency effect that informs us about the influence of flanker stimuli in a similar way as the congruency effects found in reaction time and accuracy, but now this is reflected in a single measure that is easier to interpret. Three experiments were carried out to examine whether time pressure indeed results in tunnel vision. Different task settings with varying interstimulus distances were employed as a variation in the distance between target and flankers might play an important role in the observed effects.

## II. GENERAL METHOD

### A. Overview and Apparatus

Three experiments were performed in which the same method was used. In these experiments, participants were seated in front of a 17" color CRT monitor at approximately 0.8 m viewing distance. Responses were given by pressing the left or right control (ctrl) key on a standard QWERTY keyboard with the corresponding index finger. Presentation software (Neurobehavioral Systems, Inc., 2012) was used for the presentation of instructions, stimuli, feedback, and for the recording of responses.

### B. Stimuli and Procedure

**Trial structure.** A red rectangle ( $10^\circ \times 1^\circ$ ) containing a white fixation cross ( $0.7^\circ \times 0.7^\circ$ ) was presented on a black background in the center of the screen at the onset of a trial. After 750 ms the fixation cross was replaced by the target arrowhead pointing to the left or to the right. Four flankers were presented simultaneously, two on each side of the target. The four flankers were identical within each trial and were pointing either in the same direction as the target (congruent condition), or the opposite direction (incongruent condition), or they were equal signs (neutral condition).

Stimuli and flankers were all  $0.7^\circ$  wide. Immediately after stimulus presentation, the color of the rectangle gradually faded from red to black, indicating the available time to respond. Feedback was provided immediately after a response or a missed deadline. The feedback consisted of a short text in Dutch which can be translated as „Correct”, „Incorrect” or „Too late”. In Figure 2, an overview of the events on a single trial is displayed. For incorrect and late responses, the text was accompanied by a loud 'buzzer' sound. The duration of the feedback was dependent on the duration of stimulus presentation, so that the total trial duration could be kept constant at 2500 ms.

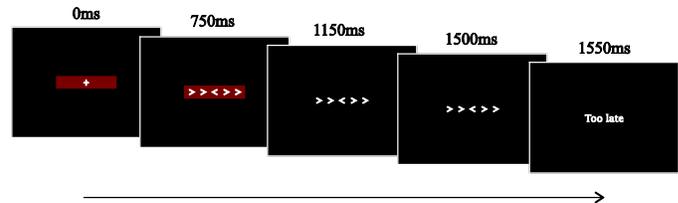


Fig. 2 The structure of a trial. An example of the various displays on a trial with a small interstimulus distance, incongruent flankers, low time pressure, and no response given before the deadline. Note that the color of the rectangle gradually changes from array onset at 750 ms till the response deadline at 1550 ms.

In the low time pressure condition, the response deadline was set at 800 ms after stimulus onset. At that moment the red background rectangle had become totally black. In the high time pressure condition the deadline, the moment at which the rectangle turned black, varied based on ongoing performance in order to keep time pressure on a relatively high level. At the start of a high time pressure block, the deadline was set at 450 ms. The available response time was reduced after two consecutive correct and fast-enough trials. After every incorrect or too slow response, the available time was increased. The initial step size for adjusting the deadline was set at 60 ms. After the first change in adjustment direction, the step size was reduced to 15 ms.

**Procedure.** A session began with a short oral introduction by the experimenter, followed by written instructions presented on the monitor. One very slow practice trial (with a 2000 ms deadline) was then presented. Next, a short instruction announced the start of a practice block of ten trials, indicating that responses had to be made faster as compared to the first trial as signaled by the faster color fading of the rectangle. After the practice block, the participant was asked if the task was clear. When the participant indicated to be ready, the experimenter left the experimental room, and the participant began with the first experimental block.

The experimental session consisted of eight blocks with a mandatory five minute break between the fourth and the fifth block. Low and high time pressure blocks alternated, but the session always started with a low time pressure block. Before each block, a short instruction was presented on the screen. The instructions preceding a low time pressure block stated that the response deadline was constant throughout the block. The instructions preceding a high time pressure block stated that the response deadline varied per trial.

A block consisted of 44 congruent trials, 44 incongruent trials and 22 neutral trials, with an equal number of left and right

targets in each condition, resulting in a total of 110 trials per block. The trials within a block were presented in random order with the restriction that the same stimulus array was not repeated on more than three consecutive trials.

The first ten trials of each block were additionally regarded as practice trials, to enable the subject to adjust to the time pressure level of the block. Responses with the incorrect hand, premature responses ( $RT < 150$  ms) and too slow responses ( $RT > 800$  ms) were defined as errors. Note that responses made with the correct hand after the deadline but before 800 ms in the high time pressure condition resulted in negative feedback („too slow”) but these responses were not treated as errors in the behavioral analyses.

The mean RT of correct responses and the mean PC was calculated for each participant in each of the experimental conditions. Mean RTs and PCs were submitted to an analysis of variance (ANOVA) for repeated measures. Greenhouse-Geisser  $\epsilon$  correction was applied whenever appropriate. Significant effects were further examined using  $t$ -tests.

### C. Hierarchical Drift Diffusion Model (HDDM)

We used the hierarchical version of the drift diffusion model developed by [6] called the HDDM. The HDDM allows the inclusion of all observed data (responses and reaction times) from all conditions and all participants in a joint analysis. In the HDDM, effects are allowed to vary over participants and conditions, enabling the analysis of multiple effects in one simulation. Following the notation of [6], we used indices to indicate the levels of differentiation and defined the Wiener distribution as follows:

$$Y_{(phij)} \sim Wiener(\alpha_{(ph)}, \beta, \tau_{(phij)}, \delta_{(phij)}),$$

The index  $p$  represents a participant,  $h$  a time pressure condition,  $i$  a congruency condition, and  $j$  an individual trial. The indices indicate that the value of the boundary  $\alpha$  can vary across persons and across time pressure conditions, the value of  $\beta$  is invariant, and  $\tau$  can differ across participants, time pressure conditions and trials. At the second level, for each time pressure condition, the boundary separation parameters  $\alpha_{(ph)}$  are assumed to be normally distributed with an inter-participant mean and variance. The non-decision time parameter  $\tau_{(phij)}$  is assumed to be normally distributed with a participant-specific mean ( $\theta_{(p)}$ ) and standard deviation ( $\chi_{(p)}$ ):  $\tau_{(phij)} \sim N(\theta_{(p)}, \chi_{(p)})$ . The participant's mean is assumed to be sampled from a normal distribution:  $\theta_{(p)} \sim N(\mu_{\theta}, \sigma_{\theta})$ . The standard deviation,  $\chi_{(p)}$ , representing the variability in decision time across participants, is assumed to be a priori uniformly distributed on a positively restricted interval, which specifies the variability in participants' standard deviations in the population. We allowed the drift rate parameter  $\delta$  to differ on each trial, and assumed it is normally distributed with an intertrial mean  $\delta_{(phij)} \sim N(v_{(phi)}, \eta_{(p)})$ . We further assumed that this participant-specific mean  $v$  is distributed according to an inter-participant normal distribution that differs across time pressure condition and experimental condition according to  $v_{(phi)} \sim N(\mu_{v(hi)}, \sigma_{v(hi)})$ . The standard deviation of  $\delta$  differs across participants, and is also uniformly distributed on a positively restricted interval. A graphical representation of this model and its assumptions is depicted in Figure 3.

In this model, the indices of  $\alpha$  are  $p$  and  $h$ , which indicates that a value of  $\alpha$  is defined for both time pressure conditions for each individual participant. The indices of drift rate include not only  $p$  and  $h$ , but also  $i$  and  $j$ , which indicates that a value for  $\delta$  is estimated for every individual trial. Since we do not focus on individual trials, but on the general effects of time pressure and congruency, we will use its inter-trial mean  $v_{(phi)}$  in our analysis of the drift rate.

Response data from the experiment were transformed in preparation of model parameter fitting. Most notably, RTs for both correct and incorrect responses were included, with RTs for incorrect responses being negated to distinguish them from correct responses.

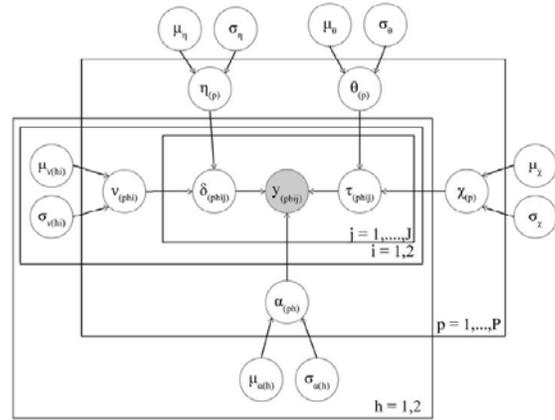


Fig. 3 A graphical representation of the hierarchical model used in our experiments. The shaded node  $y_{phij}$  represents the observed data (reaction times of correct and incorrect trials). The nodes  $\alpha_{ph}$ ,  $\delta_{phij}$ , and  $\tau_{phij}$  represent the main parameters of the model. The index  $h$  indicates the time pressure condition, and has two possible values. The index  $i$  represents the experimental condition. It has three possible values in Experiment 1 and 2 (representing the three target-flanker congruency levels), and six possible values in Experiment 3, where congruency and interstimulus distance are combined. The index  $p$  represents a participant, and the index  $j$  represents a trial. The boundary separation  $\alpha_{ph}$  is allowed to vary between participants and time pressure condition, and is assumed to be normally distributed with an interparticipant mean  $\mu_{\alpha(h)}$  and variance  $\sigma_{\alpha(h)}$ . The drift rate  $\delta_{phij}$  can vary between individual trials, but is assumed to be normally distributed with an intertrial mean  $v_{phi}$  and variance  $\eta_p$ . The intertrial mean drift rate,  $v_{phi}$ , is assumed to be normally distributed with mean  $\mu_{v(hi)}$  and variance  $\sigma_{v(hi)}$ . The non-decision time  $\tau_{phij}$  can vary between trials, but is assumed to be normally distributed with a participant specific mean,  $\theta_p$ , and variance,  $\chi_p$ . Vague uniform priors were specified for the prior parameters  $\eta_p$ ,  $\theta_p$ , and  $\chi_p$ .

The model definition and prepared data are used as input for a model parameter fitting process using software developed by [6]. This software uses Bayesian statistical methods to estimate parameter values. Two separate simulations, called chains, using the same model and the same data but different starting values for all parameters, were run for 10,000

iterations. After this initial part of the simulation, convergence of both chains is checked using visual inspection and the Gelman-Rubin statistic for all parameters of interest. Convergence is reached if the original starting values of the

estimated parameters have no influence on the current estimates. This was evaluated by comparing the values of the chains with different starting values. When we were satisfied that convergence had been met, another 30,000 iterations were

Table 1. Mean Reaction Times (RT) and Percentage of Correct Responses (PC) in Low and High Time Pressure Conditions.

| Distance     | Time Pressure | RT        |         |             | PC        |         |             |
|--------------|---------------|-----------|---------|-------------|-----------|---------|-------------|
|              |               | Congruent | Neutral | Incongruent | Congruent | Neutral | Incongruent |
| Experiment 1 |               |           |         |             |           |         |             |
| 1.4°         | Low           | 398       | 404     | 436         | 97.9      | 97.8    | 91.6        |
|              | High          | 361       | 366     | 386         | 92.7      | 90.4    | 76.0        |
| Experiment 2 |               |           |         |             |           |         |             |
| 3.5°         | Low           | 404       | 407     | 411         | 96.1      | 95.8    | 94.6        |
|              | High          | 362       | 368     | 368         | 84.4      | 84.0    | 81.5        |
| Experiment 3 |               |           |         |             |           |         |             |
| 1.4°         | Low           | 396       | 402     | 419         | 96.4      | 94.4    | 94.1        |
|              | High          | 344       | 348     | 357         | 85.2      | 81.9    | 79.9        |
| 3.5°         | Low           | 395       | 400     | 406         | 97.6      | 91.1    | 96.6        |
|              | High          | 346       | 347     | 351         | 86.6      | 84.1    | 74.1        |

run to create the posterior distribution. The results of the analysis are the posterior probability distributions of the parameters, which describes the estimated value and confidence interval after having observed the data.

The estimated mean parameter values for the response criterion  $\alpha$ , and the drift rate  $\nu$  for each participant and each experimental condition were further analyzed using a repeated measures ANOVA.

### III. EXPERIMENT 1

#### A. Method

**Participants.** Eighteen students (mean age 21 years, 12 females, 1 left-handed) with normal or corrected-to-normal visual acuity participated in this experiment. All participants signed an informed consent form and received course credits for their participation. The experiment was approved by the ethics committee of the Faculty of Behavioral Sciences at the University of Twente.

#### B. Stimuli and Procedure.

The stimuli and procedure used are described in the General Method. In this experiment the interstimulus distance was set at 1.4°.

**Data Analysis.** The model as defined before  $Y_{(phij)} \sim \text{Wiener}(\alpha_{(ph)}, \beta, \tau_{(phij)}, \delta_{(phij)})$  was used with indices  $p$  for participants ( $p=1, \dots, P$ ),  $h$  for time pressure ( $h=1, 2$ ),  $i$  for congruency condition ( $i=1, 2, 3$ ), and  $j$  for trial ( $j=1, \dots, J$ ). Where  $P$  = the number of participants (18) and  $J$  = the total number of included trials (14,352). We assumed no bias:  $\beta=0.5$ . After the first 10,000 iterations all parameters of interest had a Gelman Rubin statistic under 1.1. Visual inspection of the two chains showed no signs of convergence problems.

**Results.** After dismissing the first ten trials of each block as training trials, a total of 14,400 trials remained for the analyses. Of those trials, only three trials had premature responses ( $RT < 150\text{ms}$ ), 45 trials had too late responses ( $RT > 800\text{ms}$ ) or no response, and 1,320 trials had erroneous responses. The mean RTs and PCs for each combination of time pressure and congruency condition are shown in the upper panel of Table 1.

Mean RT for each condition was calculated for each participant and submitted to an ANOVA for repeated measures. Participants responded faster in the high time pressure than in the low time pressure condition,  $F(1,17) = 141.2, p < 0.001, \eta^2_{\text{partial}} = 0.89$ , indicating the effectiveness of our time pressure manipulation. The standard effect of flanker congruency was also observed,  $F(2,34) = 48.9, p < 0.001, \varepsilon = 0.54, \eta^2_{\text{partial}} = 0.74$ , with fastest responses in the case of congruent flankers and slowest responses in the case of incongruent flankers.

Our main interest concerned the possible interaction between time pressure and congruency. Tunnel vision was thought to result in a decreased congruency effect in the high time pressure condition. An interaction between time pressure and congruency was indeed observed,  $F(2,34) = 13.5, p < 0.001, \varepsilon = 0.74, \eta^2_{\text{partial}} = 0.44$ . The congruency effect was significantly smaller in the high time pressure condition ( $M=25, SD=17$ ) as compared to the low time pressure condition ( $M=37, SD=19$ ),  $t(17) = 4.66, p < 0.001$ . This reduced effect of target-flanker congruency under high time pressure might be an indication of tunnel vision.

Mean PC for each condition was calculated for each participant and submitted to an ANOVA for repeated

measures. Participants were less accurate in the high time pressure as compared to the low time pressure condition  $F(1,17) = 162.1, p < 0.001, \eta^2_{partial} = 0.91$ . The congruency effect was also present, with the highest accuracy on

congruent trials, and the lowest accuracy on incongruent trials,  $F(2,34) = 29.7, p < 0.001, \varepsilon = 0.55, \eta^2_{partial} = 0.64$ . As with RT, a significant interaction between time pressure and congruency was found,  $F(2,34) = 25.9, p < 0.001, \varepsilon =$

Table 2. Estimated Values for the Response Criterion ( $\alpha$ ) and the Drift Rate ( $\nu$ ), in Low and High Time Pressure Conditions.

| Distance     | Time Pressure | $\alpha$ |           |         | $\nu$       |  |  |
|--------------|---------------|----------|-----------|---------|-------------|--|--|
|              |               |          | Congruent | Neutral | Incongruent |  |  |
| Experiment 1 |               |          |           |         |             |  |  |
| 1.4°         | Low           | 0.0856   | 0.6138    | 0.5786  | 0.3763      |  |  |
|              | High          | 0.0443   | 0.5966    | 0.5313  | 0.2680      |  |  |
| Experiment 2 |               |          |           |         |             |  |  |
| 3.5°         | Low           | 0.0787   | 0.5143    | 0.4899  | 0.4551      |  |  |
|              | High          | 0.0398   | 0.4561    | 0.4223  | 0.3935      |  |  |
| Experiment 3 |               |          |           |         |             |  |  |
| 1.4°         | Low           | 0.0866   | 0.5067    | 0.4789  | 0.3640      |  |  |
|              | High          | 0.0394   | 0.4981    | 0.4274  | 0.2733      |  |  |
| 3.5°         | Low           | 0.0866   | 0.5038    | 0.4640  | 0.4313      |  |  |
|              | High          | 0.0394   | 0.4611    | 0.4045  | 0.3684      |  |  |

0.63,  $\eta^2_{partial} = 0.60$ . However, in this case the congruency effect was larger in the high time pressure condition ( $M=16.8, SD=9.8$ ) as compared with the low time pressure condition ( $M=6.4, SD=9.6$ ),  $t(17) = 6.0, p < 0.001$ . Thus, in contrast with the RT findings, the PC data suggest an increased influence of flankers under high time pressure.

Together the PC and RT data cannot answer the question whether time pressure induced tunnel vision as increased time pressure resulted in a decreased influence of flankers on RT, but an increased influence on PC. Examination of the parameters estimated with the HDDM may help in understanding the influence of time pressure.

**HDDM parameter estimates.** After the first 10,000 iterations, convergence was checked and these iterations were discarded. The results of the remaining iterations were used to calculate the mean estimated values of the variables of interest. Table 2 shows the estimated means for the relevant parameters. The value of  $\alpha$  represents the response criterion, where a higher value of  $\alpha$  indicates a higher response criterion (i.e., a more conservative strategy). The time pressure manipulation resulted in a reduction of 48% of the response criterion, which was highly significant,  $t(17) = 11.0, p < 0.001$ . Thus, according to the model, the required evidence for a decision was largely reduced in the case of high time pressure.

The value of  $\nu$  represents the drift rate or the rate of evidence accumulation. The mean values of  $\nu$  depicted in Table 2 show

three effects. First, a congruency effect was observed; the drift rate was highest on congruent trials, intermediate on neutral trials, and lowest on incongruent trials  $F(2,34) = 118.5, p < 0.001, \varepsilon = 0.54, \eta^2_{partial} = 0.88$ . Second,  $\nu$  was smaller in the high time pressure than in the low time pressure condition, indicating a decrease in the drift rate under high time pressure,  $F(1, 17) = 37.5, p < 0.001, \eta^2_{partial} = 0.69$ . Third, the difference between congruent and incongruent trials was larger in the high time pressure condition ( $0.596-0.268 = 0.328$ ) than in the low time pressure condition ( $0.614-0.376 = 0.238$ ),  $F(2,34) = 13.4, p = 0.001, \varepsilon = 0.65, \eta^2_{partial} = 0.44$ . This observation suggests that flankers had a larger influence on the drift rate in the case of high time pressure than in the case of low time pressure.

### C. Discussion

Time pressure resulted in faster but less accurate responses, indicating the presence of speed-accuracy trade-off. Responses on congruent trials were faster and more accurate than responses on neutral trials, and responses on incongruent trials were slowest and the least accurate. Thus, as demonstrated in numerous studies with the Eriksen task, the irrelevant flanker stimuli clearly affected performance. In the high time pressure condition, the congruency effect on RT was smaller than in the low time pressure condition. The congruency effect found on PC, however, was larger in the high time pressure condition than in the low time pressure condition. The RT data thus suggest a decreased influence of flankers under high time pressure, while the PC data suggest an opposite, namely increased effect of flankers. These results indicate that it is not possible to conclude that increased time pressure led to a decreased effect of flankers, which might be expected to occur in the case of tunnel vision.

Values for the parameters describing the underlying decision process according to the HDDM were estimated to provide insight in the observed effects. In the high time pressure condition, the value of  $\alpha$  was much smaller than in the low time pressure condition, indicating that less evidence had to be accumulated before a decision was made. Thus, time pressure induced a lowering of the response criterion. The value of the drift rate  $\nu$  was influenced by target-flanker congruency. As expected, evidence accumulated faster on congruent trials as compared to incongruent trials, indicating an influence of the task-irrelevant flankers. The drift rate, however, was also reduced in the case of high time pressure, suggesting that evidence accumulation was slowed down. This seems counterintuitive, as one might hypothesize that in the case of high time pressure extra attentional resources are allocated to

the target leading to a higher drift rate by a mechanism denoted as gain modulation (e.g., see [14]). This issue will be addressed in the General Discussion. Importantly, the difference in drift rate between congruent and incongruent trials was estimated to be larger in the case of high time pressure than in the case of low time pressure, which seems mainly due to a decrease in the drift rate on incongruent trials in the case of high time pressure. This suggests that incongruent flankers had a larger negative effect on the accumulation of information in that condition. This pattern of results is completely opposite to the predicted effect in the case of tunnel vision. We hypothesized that if tunnel vision was induced, it should result in less processing of the flanker stimuli, and therefore a smaller congruency effect. Thus, no support was obtained for the view that time pressure induces tunnel vision.

In this first experiment, the interstimulus distance was set at  $1.4^\circ$ , resulting in a strong congruency effect. This relatively small distance could possibly explain the absence of evidence for tunnel vision. It may be that attention was more focused under high time pressure, but not sufficiently so to exclude processing of the flanker stimuli. To investigate this possibility, we increased the interstimulus distance to  $3.5^\circ$  in our second experiment.

#### IV. EXPERIMENT 2

##### A. Method

**Participants.** Twenty students (mean age 19.5 years, 15 females, 2 left-handed) with reported normal or corrected-to-normal visual acuity participated in this experiment. All participants signed an informed consent form and received course credits for their participation. The experiment was approved by the ethics committee of the Faculty of Behavioral Sciences at the University of Twente.

##### B. Stimuli, Procedure and Data Analysis.

The interstimulus distance was set at  $3.5^\circ$ . The width of the background rectangle was increased accordingly to fit the complete stimulus array.

**Results.** A total of 16,000 trials remained for analysis after dismissing the first ten trials of each block as practice trials. Twenty-three trials had premature responses ( $RT < 150\text{ms}$ ), 63 trials had too late responses ( $RT > 800\text{ms}$ ) or no response, and 1,627 trials had erroneous responses.

Table 1 shows mean RT and PC for each condition. A repeated measures ANOVA revealed that time pressure resulted in faster,  $F(1,19) = 119.6, p < 0.001, \eta^2_{\text{partial}} = 0.86$ , and less accurate responses,  $F(1,19) = 91.7, p < 0.001, \eta^2_{\text{partial}} = 0.83$ . A small but significant congruency effect was found, with slower,  $F(2,38) = 10.4, p < 0.001, \varepsilon = 0.97, \eta^2_{\text{partial}} = 0.35$  and less accurate responses on incongruent as compared to congruent trials,  $F(2,38) = 6.6, p = 0.003, \varepsilon = 0.96, \eta^2_{\text{partial}} = 0.26$ . There was no significant interaction between time pressure and congruency, neither for RT,  $F(2,38) = 1.6, p = 0.22, \varepsilon = 0.70, \eta^2_{\text{partial}} = 0.08$ , nor for PC,  $F(2,38) = 0.88, p = 0.43, \varepsilon = 0.89, \eta^2_{\text{partial}} = 0.04$ . Pairwise comparisons additionally revealed that responses on incongruent trials were significantly slower,  $t(19) > 3.6, p < 0.002$ , and less accurate,

$t(19) > 2.3, p < 0.037$ , as compared to responses on congruent trials, both in the high and the low time pressure conditions. Responses on neutral trials ( $M=367, SD=29$ ) were slower as compared to responses on congruent trials ( $M=360, SD=30$ ) but only in the high time pressure condition,  $t(19) = 3.3, p = 0.004$ . Responses on neutral trials ( $M=84, SD=6$ ) only differed significantly,  $t(19) = 2.3, p = 0.034$ , on PC with responses on incongruent trials ( $M=81, SD=7$ ) in the high time pressure condition. All other differences between the neutral condition and the congruent or incongruent condition were not significant. Differences in RT between the low and high time pressure conditions were significant for all congruency conditions,  $t(19) > 8.7, p < 0.001$ , this was also the case for PC,  $t(19) > 8.4, p < 0.001$ .

**HDDM parameter estimates.** After the first 10,000 iterations, convergence was checked and these iterations were discarded as burn-in. Table 2 shows the mean estimated parameter values for  $\alpha$  and  $\nu$ . The values for  $\alpha$  show a similar pattern as in the first experiment, with a 49% reduction of the response criterion in the high time pressure condition,  $t(19) = 11.2, p < 0.001$ . The estimated means of  $\nu$  were submitted to a repeated measures ANOVA to examine the effect of flanker congruency and time pressure. As in the first experiment, the expected effect of flanker congruency was observed,  $F(2,38) = 33.9, p < 0.001, \varepsilon = 0.96, \eta^2_{\text{partial}} = 0.64$ , with the highest drift rate on congruent trials and the lowest drift rate on incongruent trials. The effect of time pressure on  $\nu$  was also replicated,  $F(1,19) = 31.1, p < 0.001, \eta^2_{\text{partial}} = 0.62$ , showing a reduction of the drift rate in the case of high as compared to low time pressure. No interaction effect between time pressure and congruency was observed,  $F(2,38) = 0.4, p = 0.67, \varepsilon = 0.91, \eta^2_{\text{partial}} = 0.02$ .

##### C. Discussion

Increased time pressure resulted in faster and less accurate responses, revealing again a tradeoff between speed and accuracy. The influence of the flanker stimuli on both speed and accuracy was small, but still significant. In contrast with our first experiment, time pressure did no longer affect the influence of flanker stimuli on RT and PC.

The effect of time pressure on the response criterion  $\alpha$  was again clearly present, which suggests that less evidence was needed to emit a response in the case of high time pressure. The congruency effect was again reflected in the estimated value of the drift rate  $\nu$ , with the same pattern of results as in our first experiment, but the size of the effect was much smaller. In contrast with the first experiment, no interaction between time pressure and congruency was observed for  $\nu$ . A main effect of time pressure on  $\nu$  was present, with again a smaller drift rate in the case of high time pressure as compared to low time pressure.

Together, the HDDM estimates of both experiments revealed a decrease of the response criterion and a decrease of the drift rate due to increased time pressure. Furthermore, the drift rate slows down on incongruent as compared to neutral and congruent trials. The major difference between the two experiments concerns the presence of an interaction between time pressure and congruency on  $\nu$  in our first experiment and

the absence of this effect in our second experiment. This difference seems likely due to the increased interstimulus distance. These results may be explained by precisely the opposite mechanism as tunnel vision, namely, a reduction in the efficiency of attentional allocation in the case of high time pressure. This may result in an increased flanker effect under high time pressure in Experiment 1, while the interstimulus difference in Experiment 2 may have been too large to exert an increased flanker effect.

An alternative possibility to be considered is the presence of different strategies in both experiments. To examine this, a third experiment was carried out in which interstimulus distance was varied. Trials with a small interstimulus distance and trials with a large interstimulus distance were randomly intermixed within all blocks, which will discourage the employment of different strategies. If the pattern of results found in the first two experiments is replicated then it seems that the observed differences were not due to the application of different strategies.

## V. EXPERIMENT 3

### A. Method

**Participants.** Seventeen students (mean age 22 years, 11 females, 1 left-handed) with reported normal or corrected-to-normal visual acuity participated in this experiment. All participants signed an informed consent form and received course credits for their participation. The experiment was approved by the ethics committee of the Faculty of Behavioral Sciences at the University of Twente.

### B. Stimuli and Procedure.

The third experiment combines the previous experiments: the general method is the same but interstimulus distance was added as a within-subject variable. A block of trials now contained 110 trials with a small interstimulus distance and 110 trials with a large interstimulus distance, and these trials were randomly intermixed within each block. The number of blocks was the same as in Experiment 1 and 2.

**Data Analysis.** The same HDDM was used as in the first two experiments; no extra hierarchical level was added to include the factor interstimulus distance. Instead the combination of congruency and interstimulus distance implied that there were simply more levels of the stimulus factor. Thus, the structure of the HDDM remained the same, but the index  $i$  had now six instead of three possible values. After parameter fitting, mean parameter values for congruency and interstimulus distance were derived from the estimated parameters values for each stimulus condition. Interstimulus distance was included as an independent variable in the ANOVAs of  $\alpha$  and  $\nu$ .

**Results.** As in the previous experiments, the first ten trials of each block were dismissed as they were considered as practice trials. Of the remaining 28,560 trials, 291 trials had premature responses (RT < 150ms), 277 trials had too late responses (RT > 800ms) or no response, and 3,309 trials had erroneous responses. The mean RT and PC for each condition are depicted in the lower panel of Table 1. Time pressure resulted in faster,  $F(1,16) = 164.3$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = 0.91$ , and less accurate responses,  $F(1,16) = 133.9$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = 0.89$ .

The congruency effect was also present on both RT,  $F(2,32) = 45.1$ ,  $p < 0.001$ ,  $\varepsilon = 0.73$ ,  $\eta^2_{\text{partial}} = 0.74$ , and PC,  $F(2,32) = 22.8$ ,  $p < 0.001$ ,  $\varepsilon = 0.79$ ,  $\eta^2_{\text{partial}} = 0.59$ , with fastest and most accurate responses on congruent trials, and slowest and least accurate responses on incongruent trials. Time pressure interacted with congruency on RT,  $F(2,32) = 9.0$ ,  $p = 0.001$ ,  $\varepsilon = 0.88$ ,  $\eta^2_{\text{partial}} = 0.36$ , and on PC,  $F(2,32) = 18.3$ ,  $p < 0.001$ ,  $\varepsilon = 0.72$ ,  $\eta^2_{\text{partial}} = 0.53$ , showing the strongest effect of time pressure on incongruent trials, both in a reduction of RT and a reduction of accuracy. Interstimulus distance had a main effect on RT,  $F(1,16) = 10.9$ ,  $p = 0.005$ ,  $\eta^2_{\text{partial}} = 0.40$ , with faster responses in the case of the largest interstimulus distance, but no effect was present on PC,  $F(1,16) = 0.42$ ,  $p = 0.526$ ,  $\eta^2_{\text{partial}} = 0.03$ . The interstimulus distance interacted with congruency on RT,  $F(2,32) = 17.0$ ,  $p < 0.001$ ,  $\varepsilon = 0.91$ ,  $\eta^2_{\text{partial}} = 0.52$ , but not on PC,  $F(2,32) = 3.2$ ,  $p = 0.054$ ,  $\varepsilon = 0.95$ ,  $\eta^2_{\text{partial}} = 0.17$ . No interaction was found between interstimulus distance and time pressure on RT,  $F(1,16) = 3.7$ ,  $p = 0.073$ ,  $\eta^2_{\text{partial}} = 0.19$ , and also not on PC,  $F(1,16) = 1.3$ ,  $p = 0.267$ ,  $\eta^2_{\text{partial}} = 0.08$ . The interaction between interstimulus distance, time pressure and congruency was significant for PC,  $F(2,32) = 15.0$ ,  $p < 0.001$ ,  $\varepsilon = 0.80$ ,  $\eta^2_{\text{partial}} = 0.48$ , but not for RT,  $F(2,32) = 1.1$ ,  $p = 0.332$ ,  $\varepsilon = 0.92$ ,  $\eta^2_{\text{partial}} = 0.07$ . Separate analyses for both interstimulus distances were performed to enable a direct comparison with the results of Experiment 1 and 2.

For trials with a small interstimulus distance, time pressure reduced both RT,  $F(1,16) = 153.2$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = 0.91$ , and PC,  $F(1,16) = 120.1$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = 0.88$ . The congruency effect was also found on both RT,  $F(2,32) = 43.3$ ,  $p < 0.001$ ,  $\varepsilon = 0.73$ ,  $\eta^2_{\text{partial}} = 0.73$ , and PC,  $F(2,32) = 7.5$ ,  $p = 0.002$ ,  $\varepsilon = 0.86$ ,  $\eta^2_{\text{partial}} = 0.32$ . The reduction of the congruency effect on RT under high time pressure was also replicated,  $F(2,32) = 5.5$ ,  $p = 0.009$ ,  $\varepsilon = 0.97$ ,  $\eta^2_{\text{partial}} = 0.25$ . The congruency effect on PC under high time pressure, however, was not significant,  $F(2,32) = 1.4$ ,  $p = .27$ ,  $\varepsilon = 0.88$ ,  $\eta^2_{\text{partial}} = 0.08$ , which contrasts with the results of Experiment 1.

For trials with a large interstimulus distance, time pressure again resulted in faster,  $F(1,16) = 166.9$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = 0.91$ , and less accurate responses,  $F(1,16) = 121.8$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = 0.88$ . The congruency effect was also replicated in both RT,  $F(2,32) = 18.9$ ,  $p < 0.001$ ,  $\varepsilon = 0.89$ ,  $\eta^2_{\text{partial}} = 0.54$ , and PC,  $F(2,32) = 23.8$ ,  $p < 0.001$ ,  $\varepsilon = 0.83$ ,  $\eta^2_{\text{partial}} = 0.60$ . Here, time pressure reduced the congruency effect on RT,  $F(2,32) = 5.2$ ,  $p = 0.011$ ,  $\varepsilon = 0.97$ ,  $\eta^2_{\text{partial}} = 0.24$ , and increased the congruency effect on PC,  $F(2,32) = 27.9$ ,  $p < 0.001$ ,  $\varepsilon = 0.75$ ,  $\eta^2_{\text{partial}} = 0.64$ , while such interactions were not observed in Experiment 2.

**HDDM Parameter Estimates.** Table 2 shows the mean estimated parameter values for  $\alpha$  and  $\nu$  calculated from the posterior distribution after discarding the first 10,000 iterations to ensure that convergence had been met. The estimated values for  $\alpha$  show a similar pattern as for Experiment 1 and 2, with a reduction of required evidence of 55% in the case of high time pressure relative to low time pressure,  $t(16) = 10.1$ ,  $p < 0.001$ .

A repeated measures ANOVA on the estimated values of  $\nu$  with the factors interstimulus distance, congruency, and time

pressure revealed the following results. The congruency effect,  $F(2,32) = 74.0$ ,  $p < 0.001$ ,  $\varepsilon = 0.62$ ,  $\eta^2_{\text{partial}} = 0.82$ , and the effect of time pressure were replicated,  $F(1,16) = 9.8$ ,  $p = 0.006$ ,  $\eta^2_{\text{partial}} = 0.38$ . An interaction between time pressure and congruency was found,  $F(2,32) = 9.5$ ,  $p = 0.001$ ,  $\varepsilon = 0.97$ ,  $\eta^2_{\text{partial}} = 0.37$ . A main effect of interstimulus distance was observed on  $v$ ,  $F(1,16) = 6.0$ ,  $p = 0.026$ ,  $\eta^2_{\text{partial}} = 0.27$ , which reflected an overall higher drift rate with the large as compared to the small interstimulus distance. Interstimulus distance also modulated the interaction between time pressure and congruency  $F(2, 32) = 5.0$ ,  $p = 0.013$ ,  $\varepsilon = 0.92$ ,  $\eta^2_{\text{partial}} = 0.24$ . Separate analyses for both distances showed that the interaction between time pressure and congruency on  $v$  was significant with the small interstimulus distance,  $F(2,32) = 12.9$ ,  $p < 0.001$ ,  $\varepsilon = 0.92$ ,  $\eta^2_{\text{partial}} = 0.45$ , but not with the large interstimulus distance,  $F(2,32) = 1.2$ ,  $p = 0.326$ ,  $\varepsilon = 0.99$ ,  $\eta^2_{\text{partial}} = 0.07$ . Specifically, in the case of the small interstimulus distance, the reduction of the drift rate due to time pressure was small for congruent trials but large for incongruent trials, while no such effect was present in the case of the large interstimulus distance.

### C. Discussion

The main effects of time pressure and congruency as observed on our behavioral measures in Experiment 1 and 2 were replicated in our third experiment. A direct comparison of the results for the trials with a small interstimulus distance with the result of Experiment 1 shows a comparable pattern. However, the interaction between time pressure and congruency on PC did not reach significance in our third experiment. A direct comparison of the results for the trials with a large interstimulus distance with the results from Experiment 2 also showed some minor differences. In our third experiment, we observed a significant interaction between time pressure and congruency on both RT and PC that was not found in the second experiment. An examination of the estimated parameters for the drift rate and the response criterion might clarify whether these observed differences point to different conclusions.

Separate analyses of the estimated drift rate for both interstimulus distances revealed quite comparable effects of time pressure and congruency as in Experiment 1 and 2. Time pressure and congruency both affected the drift rate on trials with a small and large interstimulus distance. Importantly, time pressure increased the congruency effect for trials with a small interstimulus distance and did not influence the congruency effect for trials with a large interstimulus distance, which implies that the results on the drift rate as observed in Experiment 1 and 2 were replicated in our third experiment.

## VI. GENERAL DISCUSSION

In this paper, the central question to be addressed was whether tunnel vision, a shrinkage in the size of the attentional focus, can be demonstrated in the case of stressful conditions. To answer this question, we employed an arrowhead-version of the Eriksen flanker task, in which a central target was accompanied by congruent, neutral, or incongruent flankers. Stress was induced by varying time pressure between conditions. Three experiments were carried out in which

different interstimulus distances were employed. Interest was focused on behavioral measures indicating possible effects of tunnel vision, and especially on parameters of the underlying decision process that can be estimated with the HDDM: the drift rate ( $v$ ), and the height of the response criterion ( $\alpha$ ). We expected to observe that increased time pressure would lead to a reduction of the response criterion, and that congruency of flankers would affect the drift rate, with the highest drift rate in the case of congruent flankers and the lowest drift rate in the case of incongruent flankers: a congruency effect. Most importantly, we reasoned that tunnel vision (in the case of high time pressure) would be reflected in a reduction of the congruency effect on the drift rate.

Behavioral results revealed clear effects of time pressure and flanker congruency in all our experiments. Responses were faster and less accurate when time pressure was high, demonstrating a speed-accuracy tradeoff. Responses were faster and more accurate in the case of congruent than in the case of incongruent flankers. However, the observed interactions between time pressure and congruency proved to be difficult to interpret as regularly opposite effects were observed on RT and PC. One of the major reasons to use the HDDM for our question of interest was to resolve this impasse.

The estimated parameters of the underlying decision process according to HDDM revealed several interesting insights. First, a consistent and expected observation was the reduction of the response criterion in the case of high as compared to low time pressure estimated on the basis of the behavioral results of our experiments. Secondly, we also consistently but unexpectedly observed a reduction of the drift rate in the case of high time pressure relative to the condition with low time pressure. Thus, time pressure reduced the response criterion but also decreased the rate of the accumulation of evidence. A possibility, considered more thoroughly below, is that the reduction of the response criterion may have been overestimated, which will thereby also affect the estimation of the drift rate. Third, we observed an interaction between time pressure and congruency on the drift rate for the conditions with the small interstimulus distance, but not for the conditions with a large interstimulus distance. Opposed to our expectations, this interaction in the case of a small interstimulus distance actually reflected a larger congruency effect in the case of high time pressure and not a reduction of the congruency effect. Thus, the influence of flankers on the accumulation of evidence was increased in the case of high time pressure, at least when interstimulus distance was not too large. These findings lead to the conclusion that time pressure did not induce tunnel vision but actually decreased the efficiency of attentional allocation, which is detrimental in the case of a small interstimulus distance but not so in the case of a larger interstimulus distance. Nevertheless, before accepting this as the conclusion of this paper it seems relevant to discuss four different issues. First, our results suggest that the presented conception of tunnel vision may simply be flawed, which may imply that we have to redefine what we precisely mean with the term tunnel

vision. Secondly, according to some authors, the interpretation of the flanker effect has to be reconsidered, which has important consequences for the drawn conclusions. Third, the HDDM parameter estimates may not have been optimal, which consequently affects the interpretation of the observed effects. Finally, the generalizability of the results from our task to real life conditions may be questionable.

At the beginning of our paper, we indicated that tunnel vision may be understood as a shrinkage in the size of the attentional focus. This view is obviously based on a very literal interpretation of tunnel vision; may it not be the case that tunnel vision should be interpreted in a less literal and more metaphorical way? Of course, this all strongly relates to our view on spatial attention; can we really interpret attention as a spotlight, a zoom lens, or a gradient that varies in size? (e.g., see [15]-[17]). For example, it is becoming clearer that there are strong similarities between spatial attention and the retrieval of information from working memory (e.g., see [18]), and it seems according to several researchers in the field obvious that attention is not meant for perception but rather for interacting with the outer world. Maybe tunnel vision is better understood as a reduced ability to process all available information rather than a reduction in the size of the attentional focus.

A highly related issue concerns the interpretation of the Eriksen flanker effect. The common interpretation of the flanker effect is that participants are not able to consistently keep their attention focused on the central target but partly divide their attention across the flankers thereby invoking the benefit with congruent flankers and a cost with incongruent flankers. Recently, [19] proposed that the flanker effect does not emerge because of a failure in selecting the target from the array, in line with the aforementioned ideas, but rather as a consequence of the effectiveness of attentional selection concerning task-relevant features (for related discussions on the flanker task in terms of different variants of diffusion models, see [20]-[22]). Target-like features are simply extracted from the whole environment and not from a single location. If we extend this idea slightly further, this selection of target-like features may directly exert an effect on the selection of actions. In our task version, this implies the activation of conflicting actions in the case of incongruent flankers. Moreover, if we consider the earlier mentioned idea that time pressure also speeds up sensory processing by gain modulation (see [14]) then we might explain the presence of an enlarged congruency effect on the drift rate in our conditions with high time pressure as there will be a stronger activation of the two conflicting actions. The absence of this effect with the larger interstimulus distance may be ascribed to a reduction in the visual acuity of the flankers. Nevertheless, some other studies referred to in our introduction provided no support for an influence of time pressure on pre-motoric processes (see [3]; [5]). Furthermore, we did not find support for an increase in the drift rate on congruent trials, but mainly a decrease of the drift rate on incongruent trials. Nevertheless, it is obvious that other ideas concerning the origin of the flanker effect and attentional selection [19] have a major

impact on the meaning of a phenomenon such as tunnel vision.

In all three experiments we noticed that time pressure not only reduced the estimated response criterion, but also reduced the estimated drift rate. The latter observation seems counterintuitive, as one might rather expect (see above) the drift rate to increase in the case of high time pressure. It may be argued that the estimation of the response criterion and the drift rate on the basis of HDDM are not completely appropriate. To evaluate this, we decided to use the estimated parameter values of Experiment 3 to reproduce the observed response data, which gives an idea of the goodness-of-fit of the obtained parameter values. Figure 4 shows the observed and predicted RT distributions for each condition for three participants, with incorrect responses flipped to the left. The gray bars represent the observed data, while the open bars represent the data generated by the estimated parameter values. The predicted data match the observed data quite well. Nevertheless, although the reconstruction of the response data suggests that the obtained parameter values are appropriate it is still possible that the HDDM overestimates the effect of time pressure on the response criterion. In all our experiments, the influence of high time pressure seems very strong, as a reduction of at least 45% was observed. If we consider the possibility that this reduction is an overestimation of the influence of time pressure (i.e., the estimate of  $\alpha$  is too small), then an appropriate fit of the data can only be obtained if the effect on the drift rate is overestimated as well (i.e., the estimate of  $v$  is too small), otherwise, the reconstructed response data should display shorter response data as compared to the originally observed data.

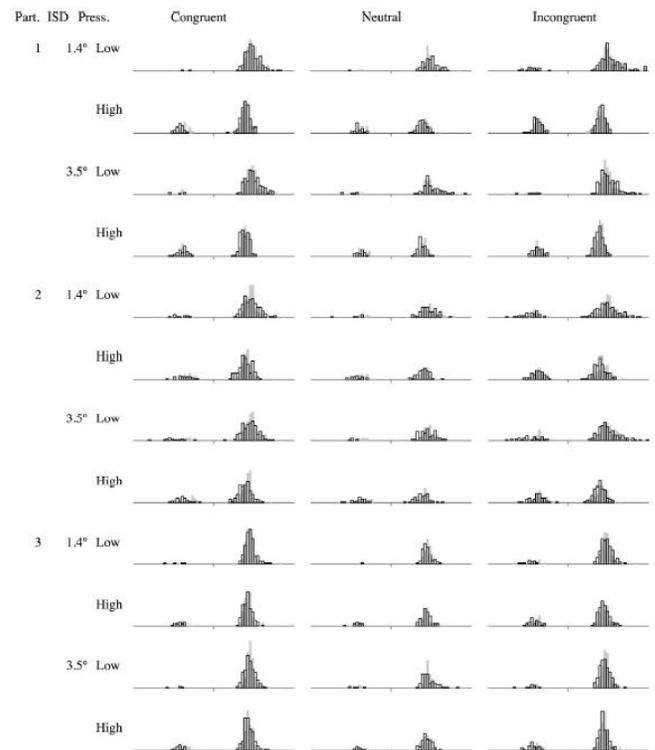


Fig. 4 Frequency histograms of observed and predicted behavioral data for the first three participants in Experiment 3. Part. =

participant, ISD = interstimulus distance, Press. = time pressure. The filled gray bars represent actual measured reaction times, divided into 50ms bins. The open bars represent the reaction times based on the estimated parameter values. The left half of each cell is a flipped histogram of the incorrect responses.

As we observed both a reduction in the response criterion and the drift rate due to high time pressure, this may very well have been the case. Moreover, earlier mentioned studies did not observe an effect of time pressure on pre-motoric processes [3],[5], which also seems not in line with an influence on the drift rate. Importantly, even though it may be the case that the reduction of the response criterion is overestimated, a possibly better estimation of the relevant parameters is likely to yield a comparable pattern of results.

Although the Eriksen flanker task used in this study seems well suited to test the presence of tunnel vision, its properties limit generalizability to other settings for several reasons. First, the stimuli that are to be ignored are always present on each trial. In real live settings such as in the car accident example in our introduction, stimuli outside the focus of attention are far from predictable. Second, in the flanker task the target is the only task-relevant stimulus while flankers are to be ignored. This categorization of stimuli as either task-relevant or task-irrelevant may also not generalize well to real-world settings as in the latter case every stimulus is potentially important.

On the basis of our behavioral data and the estimates of the underlying decision process determined by the HDDM it can be concluded that time pressure seems to lower the response criterion, while irrelevant flankers affect the speed of the accumulation of information. Opposed to our initial idea, it could not be concluded that time pressure induced tunnel vision, rather, it appears to be the case that time pressure reduced the efficiency of spatial attentional selection.

#### REFERENCES

- [1] A.E. Bursill, "The restriction of peripheral vision during exposure to hot and humid conditions," *Quarterly Journal of Experimental Psychology*, vol. 10, pp. 113-129, 1958. doi:10.1080/17470215808416265.
- [2] G.R. Dirkin, "Cognitive tunneling: Use of visual information under stress," *Perceptual and Motor Skills*, vol. 56, pp. 191-198, 1983. doi:10.2466/pms.1983.56.1.191
- [3] A. Osman, L. Lou, H. Müller-Gethmann, G. Rinkenauer, S. Mattes, and R. Ulrich, "Mechanisms of speed-accuracy tradeoff: Evidence from covert motor processes," *Biological psychology*, vol. 51, pp. 173-199, 2000. doi:10.1016/S0301-0511(99)00045-9B.
- [4] B.A. Eriksen, and C.W. Eriksen, "Effects of noise letters upon the identification of a target letter in a nonsearch task," *Perception & Psychophysics*, vol. 16, pp. 143-149, 1974. doi:10.3758/BF03203267J.
- [5] R.H.J. Van der Lubbe, P. Jaśkowski, B. Wauschkuhn, and R. Verleger, "Influence of time pressure in a simple response task, a choice-by-location task, and the Simon task," *Journal of Psychophysiology*, vol. 15, pp. 241-255, 2001. doi:10.1027//0269-8803.15.4.241
- [6] J. Vandekerckhove, F. Tuerlinckx, and M.D. Lee, "Hierarchical diffusion models for two-choice response times," *Psychological Methods*, vol. 16, pp. 44-62, 2011. doi:10.1037/a0021765.
- [7] K.R. Ridderinkhof, G.P. Band, and D. Logan, "A study of adaptive behavior: effects of age and irrelevant information on the ability to inhibit one's actions," *Acta Psychologica*, vol. 101, pp. 315-337, 1999. doi: 10.1016/s0001-6918(99)00010-4.
- [8] R.G. Pachella, and R.W. Pew, "Speed-accuracy tradeoff in reaction time: Effect of discrete criterion times," *Journal of Experimental Psychology*, vol. 76, pp. 19-24, 1968. doi:10.1037/h0021275
- [9] W.A. Wickelgren, "Speed-accuracy tradeoff and information processing dynamics," *Acta Psychologica*, vol. 41, pp. 67-85, 1977. doi:10.1016/0001-6918(77)90012-9
- [10] R. Ratcliff, "A theory of memory retrieval," *Psychological Review*, vol. 85, pp. 59-108, 1978. doi:10.1037/0033-295X.85.2.59
- [11] T.V. Wiecki, I. Sofer, and M.J. Frank, "HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python," *Frontiers in Neuroinformatics*, vol. 7, pp. a14, 2013. doi:10.3389/fninf.2013.00014.
- [12] R. Ratcliff, and J.N. Rouder, "A diffusion model account of masking in two-choice letter identification," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, pp. 127-140, 2000. doi:10.1037/0096-1523.26.1.127.
- [13] J. Zhang, and J.B. Rowe, "Dissociable mechanisms of speed-accuracy tradeoff during visual perceptual learning are revealed by a hierarchical drift-diffusion model," *Frontiers in Neuroscience*, vol. 8, pp. a69, 2014. doi:10.3389/fnins.2014.00069.
- [14] D. Standage, G. Blohm, and M.C. Dorris, "On the neural implementation of the speed-accuracy trade-off." *Frontiers in Neuroscience*, vol. 8, pp. a236, 2014. Doi:10.3389/fnins.2014.00236.
- [15] K.R. Cave, and N.P. Bichot, "Visuospatial attention: beyond a spotlight model," *Psychonomic Bulletin & Review*, vol. 6, pp. 204-23, 1999.
- [16] C.W. Eriksen, and J.D. St. James, "Visual attention within and around the field of focal attention: a zoom lens model," *Perception & Psychophysics*, vol. 40, pp. 225-40, 1986.
- [17] D. LaBerge, and V. Brown, "Theory of attentional operations in shape identification," *Psychological Review*, vol. 96, pp. 101-124, 1989. doi:10.1037/0033-295X.96.1.101.
- [18] R.H.J. Van der Lubbe, C. Bundt, and E.L. Abrahamse, "Internal and external spatial attention examined with lateralized EEG power spectra," *Brain Research*, vol. 1583, pp. 179-192, 2014. doi : 10.1016/j.brainres.2014.08.007.
- [19] S. Buetti, A. Lleras, and C.M. Moore, "The flanker effect does not reflect the processing of "task-irrelevant" stimuli: Evidence from inattention blindness," *Psychonomic Bulletin & Review*, pp. 1-7, 2014. doi:10.3758/s13423-014-0602-9
- [20] R. Hübner, M. Steinhauser, and C. Lehle, "A dual-stage two-phase model of selective attention." *Psychological Review*, vol. 117, pp. 759-784, 2010.
- [21] C.N. White, R. Ratcliff, and J.J. Starns, "Diffusion models of the flanker task: discrete versus gradual attentional selection," *Cognitive Psychology*, vol. 63, pp. 210-238, 2011.
- [22] R. Hübner, and T. Töbel, "Does attentional selectivity in the flanker task improve discretely or gradually?," *Frontiers in Psychology*, vol. 3, pp a434, 2012. Doi: 103389/fpsyg.2012.00434

# Towards Community Recommendations on Location-Based Social Networks

Chara Remoundou, Pavlos Kosmides, Konstantinos Demestichas, Ioannis Loumiotis, Evgenia Adamopoulou, and Michael Theologou

**Abstract**— With the explosive growth of social networks during the last decades, the discovery of social communities relevant to users’ interests has become an important subject of study in the research community. Social Networks provide a plethora of suggestions (pages, friends, points of interests, etc.) to their users, in order to enhance their experience with services that adapt to their needs. In this paper, we present a method for performing user-personalized community suggestions to users of location-based social networks. Users’ history data, related to former presence declarations through so-called “check-ins”, are being exploited so as to infer their interests and find relevant communities. For this purpose, points of interest are being categorized as venues reflecting a generalized form of users’ interests that may vary during the day. In the proposed method, we present a community recommendation scheme based on predicting venues by taking advantage of users’ and their friends’ history. The dataset we used was based on input from a well-known Location-Based Social Network. For the evaluation of our approach we use two machine-learning techniques, whose performance is compared against each other.

**Keywords**— Location-Based Social Networks; learning algorithms; community recommendations

## I. INTRODUCTION

THE expansion of social media during the last years has increased exponentially the usage of group and community activities. Users are creating and joining online communities in order to share their content and interact with users who have common interests. As a result, discovering intriguing communities, among the continuous growing number of social communities, has become a challenging task for social media users. In parallel, there has been a rapid growth of location-based social networks (LBSN), as more and more users tend to share their location during their everyday lives. These kinds of declarations, usually expressed by “check-ins”, reflect users’ main activities and interests, which can be exploited for providing user-personalised suggestions about places, friends, groups or events.

One of the most innovative features that has been introduced in social networks is personalised community recommendations. Using this feature, LBSNs can provide recommendations to users about groups and social

communities that are related to their interests. To achieve that, many researchers have used machine-learning techniques to estimate users’ interests, exploiting their former activities or social connections.

Several methods have been proposed in the literature in order to provide community recommendations to social media users. Specifically, in [1] the authors present a method for modelling communities via user-generated tags. They recommend communities to users by capturing how each community’s subject is relevant to a user’s personal tags and other community members’ tags. The authors in [2] introduce a filtering method (Combinational Collaborative Filtering) to implement community recommendations by combining semantic and user information. This method uses a hybrid training strategy that combines Gibbs sampling with the EM algorithm.

In [3], the authors suggest a soft-constraint based online Latent Dirichlet Allocation (LDA) algorithm in order to detect the latent topics across communities using the number of a user’s posts in each community. Similarly, collaborative filtering-based algorithms, graph-based algorithms, and search-based algorithms that use social tagging are also presented in [4]. In this work the authors combine and compare those methods for personalized community recommendations. Finally, in [5] the authors use the users’ preferences and relationships in order to make the proper community suggestions to a specific user. In this method the communities are being categorized by familiarity, favourability, and similarity to a user’s interests in order to perform a satisfactory recommendation.

Although a lot of researchers were involved around community recommendations on classic Social Networks, only few have focused specifically on LBSNs. For example, in [6] the authors attempt to model human activity and geographical areas by means of place categories, while they also identify user communities that visit similar categories of places.

In this paper, following the above trend, we use users’ history as well as their friends’ activities in order to estimate venues that lie into users’ interests in their daily lives. By predicting venues of interest, we can form community suggestions based on places that share coinciding events. We present a system architecture for collecting and elaborating useful data from mobile users of LBSNs, and use machine-learning techniques for providing suggestions. Specifically,

C. Remoundou, P. Kosmides, K. Demestichas, I. Loumiotis, E. Adamopoulou, M. Theologou are with the Institute of Communication and Computer Systems of the National Technical University of Athens, Iroon Polytechniou 9, 15773 Greece (phone: +30-210-7721495; e-mail: chremoundou@cn.ntua.gr).

we propose the use of a Radial Basis Function (RBF) Neural Network and compare it against a K-Means algorithm.

The rest of this paper is organized as follows. In Section II, we describe the architecture of the proposed system for collecting data from LBSNs, and for processing them using machine learning mechanisms. In Section III, we present the three different approaches that we use based on machine learning, namely K-Means and RBF Neural Network. The dataset that was employed for location prediction is presented in Section IV, while in Section V we provide the results and discuss on the performance of the three proposed algorithms. Finally, the paper is concluded in Section VI.

## II. SYSTEM DESCRIPTION

In the proposed system, users will be equipped with mobile terminals connected to a location-based social network. One of the main concerns of the proposed LBSN is to provide its users with recommendations about communities that range closely to their interests. To achieve that, we take under consideration both users' preferences as well as their friends' preferences, and provide predictions about possible venues of interest. Having these information, we can discover communities of users with the same interests, and recommend them to subscribe to one or more of them.

In Fig. 1, we present the proposed system's architecture, in which users' data are collected from mobile terminals and are sent to the cloud, where machine learning engine (MLE) mechanisms are applied in order to produce recommendations. The architecture is designed using ArchiMate® notation [7], showing the basic components and their relationship.

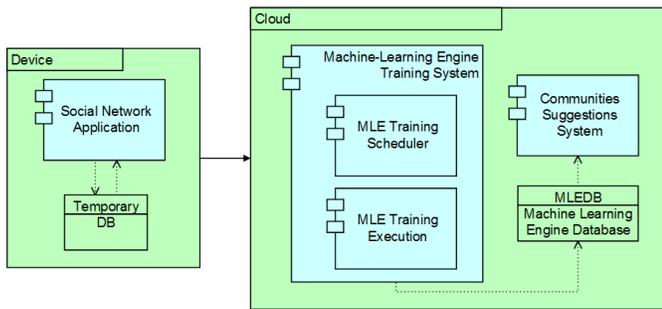


Fig. 1 System's architecture

As we may observe, the proposed system consists of one product related to the mobile device and one product representing the cloud platform.

For the mobile device, the *Social Network Application* component is responsible for collecting users' information regarding the check-ins and the ratings that users leave for each place they visit.

For the cloud platform, we recognize the following systems and components:

- Machine-Learning Engine Training System: This system is responsible for the centralized training of the machine-learning engines that will be used by the Communities Suggestions system. It is comprised of the MLE Training Scheduler and the MLE Training Execution components.

- MLE Training Scheduler: The main purpose of this component is to initiate the generation of new MLEs (such as K-Means, RBF neural networks, or PNNs).
- MLE Training Execution: This component is responsible for retrieving all necessary training data and for executing the machine learning engine training algorithm.
- Communities Suggestions System: This component is responsible for performing estimation of the venues that a user may be interested in, according to his/her own preferences, as well as the ones of his/her friends. From these venues, the appropriate communities are selected and forwarded as recommendations to users.

## III. PREDICTION MODELS

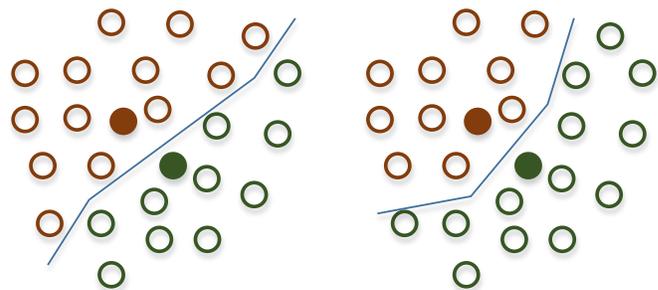
For the implementation of the proposed community recommendation system, we use two different approaches based on machine learning, namely K-Means and RBF neural network.

### A. K-Means

The first approach used as a prediction model is the K-Means Clustering. K-Means is one of the most popular clustering methods, which can be defined as the partitioning of a finite amount of data into a number of clusters by understanding the underlying structure [8].

The main idea of the K-Means algorithm is based on the following steps:

1. From an initial non optimal clustering, relocate each point to its new nearest centre.
2. Update the clustering centres by calculating the mean of the member point.
3. Repeat steps 1 and 2 until convergence criteria are satisfied (e.g. predefined number of iterations, difference on the value of the distortion function).



(a) Initial clusters

(b) Final clusters

Fig. 2 K-Means Clustering example with 2 clusters

### B. Radial Basis Function (RBF) Neural Network

RBF Neural Networks (NNs) provide an alternative to Multilayer Perceptrons. They share many features with Artificial Neural Networks (ANNs) that employ back propagation, and they are used for pattern recognition. However, RBF NNs have certain advantages which are not found in common ANNs. Specifically, the training procedures

are faster than the Multilayer Perceptrons. In addition, RBF NNs are characterised by the absence of local minima, unlike common ANNs [9].

A typical RBF NN structure, in its most basic form (Fig. 3), involves three layers, namely an input, a hidden and an output layer. In the input layer, the space can be either normalized or be an actual representation of the input data. This is then fed to the hidden layer nodes, which differ from other NNs in that each node represents a data cluster, centred at a particular point with a given radius. The hidden layer nodes are responsible for calculating the distance from the input vector to their own centre. The results are transformed using a basis function, and forwarded to the output layer. The output layer consists only one node, which sums the input from the hidden layer and produces the final result.

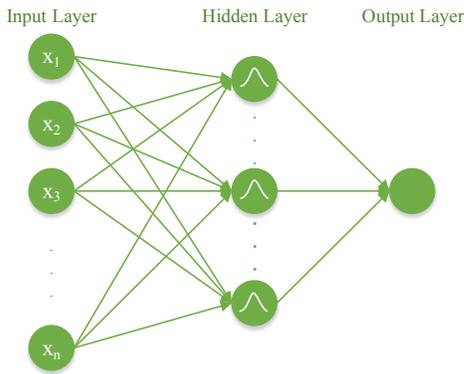


Fig. 3 Typical structure of Radial Basis Function Neural Network

#### IV. DATA ANALYSIS

The dataset used in this paper is based on measurements that were taken for research purposes in [10]. The data were collected from a well-known social network (Foursquare), which allows location-based check-ins as well as ratings.

The available files were divided according to the location that the user has visited (checkins file), the social connection between two users (socialgraph file), the venue of a specific location, e.g. restaurants (venues file), and the rating that each user has left for each location she/he has visited (ratings file).

The above mentioned files were merged, using users' id to create a seven-tuple with the variables described in Table I.

TABLE I. VARIABLES AND VALUES USED

| Variable | Description  |
|----------|--|
| user     | User's unique id {total 149 users}   |
| friend   | Unique id of user's friend   |
| day      | Values 1 – 7 representing each day of the week {Monday – Sunday}   |
| hour     | The hour that the user has checked in the specific location  |
| location | Each location is represented by a unique id created from the corresponding coordinates [(latitude, longitude)] {total 117 locations} |
| rate     | The rate that user has left for the specific location scaling 1 – 5.   |
| venue    | Unique id of venue {total 52 venues}   |

Using the described variables, the input set for the proposed algorithms can be defined as:

$$x = (user, friend, day, hour, location, rate, venue) \quad (1)$$

The restructured dataset takes into consideration not only the place that a user has visited, but also the type of each specific location (venue), as well as the rating that the user has left, depicting his/her satisfaction. We also consider users' friendships with social connections between two users, resulting to a social graph depicted in Fig. 4.

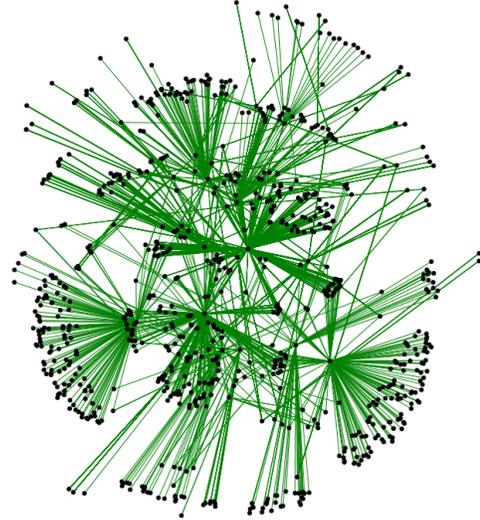


Fig. 4 Social graph using Harel-Koren Fast Multiscale lay out algorithm

#### V. RESULTS

To demonstrate the appropriateness of the proposed learning algorithms, presented in Section III, we used them in order to provide community recommendations based on users' preferences that are inferred from theirs' and their friends' history. Regarding the K-means algorithm, one of the most common parameter that we need to define, is the number of clusters that are used. In Fig. 5, we present the misclassification percentage acquired for the K-means algorithm for different numbers of clusters.

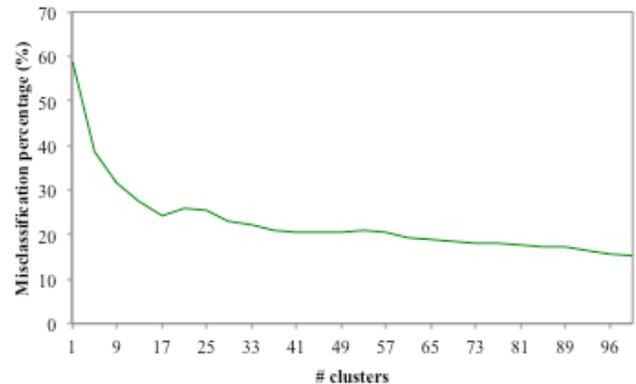


Fig. 5 Misclassification percentage for K-means algorithm with respect to number of clusters used

As we may observe, K-means performs better for an increased number of clusters. In Table II, we present the misclassification percentage that was derived for both the K-means algorithm and the RBF neural network, using 10-fold cross-validation method. Specifically, the K-means clustering was constructed with 100 clusters, while the RBF neural network was formulated with 100 neurons in the hidden layer.

TABLE II. RESULTS OF VALIDATION PROCESS

| Learning Method | Misclassification percentage (%) |
|-----------------|----------------------------------|
| K-Means         | 15.258                           |
| RBF NN          | 16.884                           |

It is clear that the K-means algorithm outperforms the RBF Neural Network, making it highly suitable for estimating venues.

## VI. CONCLUSION

In this paper, we have presented our work on providing community recommendations based on predicted venues using data from location-based social networks. Specifically, we have presented a system architecture for collecting necessary data from location-based social networks, and sending them to a cloud platform. The cloud platform is responsible for running the machine-learning algorithms in order to provide recommendations to users about possible communities that they would be interested in, where each community can be conducted from venues matching users' preferences. In order to forecast venues according to users' preferences and their social connections, we used two different learning algorithms, namely K-Means and RBF neural network. The dataset used includes information collected from a well-known social network (Foursquare), depicting users' friendships and ratings on specific locations that they have visited. The results that we received from the validation process, demonstrate that the K-Means outperforms the other proposed algorithm, giving predictions with high accuracy.

## ACKNOWLEDGMENT

This work has been performed under the Greek National project WikiZen (11ΣΥΝ\_10\_1808), which has received research funding from the Operational Programme "Competitiveness & Entrepreneurship" of the National Strategic Reference Framework NSRF 2007-2013. This paper reflects only the authors' views, and the Operational Programme is not liable for any use that may be made of the information contained therein.

## REFERENCES

- [1] A. Akther, H. Kim, M. Rawashdeh, and A. El Saddik, "LNAI 7310 - Applying Latent Semantic Analysis to Tag-Based Community Recommendations," *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, pp. 1–12, May 2012.
- [2] W.-Y. Chen, D. Zhang, and E. Y. Chang, "Combinational collaborative filtering for personalized community recommendation," *Proceedings of*

- the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 115-123, August 2008.
- [3] Y. Kang and N. Yu, "Soft-constraint based online LDA for community recommendation," *Advances in Multimedia Information Processing-PCM 2010*. Springer Berlin Heidelberg, pp. 494-505, 2011.
- [4] H.-N. Kim and A. El Saddik, "Exploring social tagging for personalized community recommendations," *User Modeling and User-Adapted Interaction*, vol. 23, no. 2–3, pp. 249–285, Sep. 2012.
- [5] H. Ko, S. Choi, and I. Ko, "A community recommendation method based on social networks for web 2.0-based IPTV," *Digital Signal Processing, 2009 16th International Conference on*. IEEE, pp. 1-6, July 2009.
- [6] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M., "Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks," In: *3rd Workshop Social Mobile Web (SMW 11)*, July 2011.
- [7] ArchiMate® 2.1 Specification, retrieved on May 2014 from URL: <http://pubs.opengroup.org/architecture/archimate2-doc/>
- [8] R.C.De. Amorim, "Constrained Intelligent K-Means: Improving Results with Limited Previous Knowledge," *The Second International Conference on Advanced Engineering Computing and Applications in Sciences*, pp. 176–180, 2008.
- [9] M. Bianchini, P. Frasconi, M. Gori, "Learning without local minima in radial basis function networks," *IEEE Transactions on Neural Networks*, vol. 6, issue 3, pp. 749-756, 1995.
- [10] J. Levandoski, M. Sarwat, A. Eldawy, and M. Mokbel, "LARS: A Location-Aware Recommender System," *ICDE Conference*, pp. 450-461, 2012.

# Numerical simulations of a pipeline crossing

Ioan Both, Adrian Ivan

**Abstract**— The paper presents the application of computer codes in the advanced analysis of structures with tension elements, cables. Static and dynamic analyses are performed for a suspension pipeline crossing by considering a real wind intensity recorded by a weather station. The effect of pre-tension is discussed from the structure Eigen-modes aspect. The static and dynamic analysis reveals different values for element forces. A construction stage analysis is performed using a particular module of finite element computer code

**Keywords**— cable, construction stage, dynamic analysis, suspension crossing.

## I. INTRODUCTION

STRUCTURAL systems supporting fluid materials transportation pipelines may be regarded as a continuous structure avoiding encountered obstacles between the two points of interest. The linear impediments such as rivers or valleys met in their path may be overcome with superstructures (above) or infrastructures (below). For both solutions advantages and disadvantages are present and for a good design the important factor is the area environment either from geometry, sustainability or protection point of view. The waterway crossing is an example that the choice of solution is influenced by several considerations. One of them refers to the environment impact. The disturbance of environment for both aquatic and terrestrial plant and animal life has to be minimized since a waterway crossing affects these factors. Both crossings, under and above water, have to take care at the hazardous and contaminated materials during construction [1]. The underwater crossing might have a greater effect over the environment because of the instability of the river bed and from here a catastrophic event is possible to occur. Waterways constructed above obstacles have the advantage that it allows better site inspections and most of the loading can be easily determined. Having a decision for a structural system over the obstacle it only remains to decide which solution of the structure is better to use: suspension

crossing, cable-stayed crossing, self-supporting or truss structure as a bearing structural system for the pipeline. The span that needs to be covered by the crossing is the main factor that influences this decision. Two reasonable solutions are suspension crossing and cable-stayed crossing. The last one has the advantage of smaller anchors and possibility of building on soft soil, but if the span increases the towers will have to increase too much [2]. To overcome this disadvantage the appropriate solution are suspension crossings.

Towers, main cables (suspension cable), hangers, anchors, lateral cables, cantilever (not necessary) and the pipeline are components of the structural system of a suspension crossing.

The load path for these structures is created as follows: the gravitational loadings from the self-weight and the loading given by the gas in the pipe is transmitted by the hangers to the main cable. The hangers have various lengths corresponding to the sag of the main cable. The main cables transmit a part of the vertical component of the force in the cable to the towers and the horizontal component is transmitted to the anchors. The horizontal actions, perpendicular to the crossing, are taken by the lateral cables (wind guy cable), connected to the pipeline by the wind ties [3].

The cable elements in these structures have an important role and it is characterized by high resistance, high flexibility, and a very small damping. Due to large displacements, suspension crossings design should consider both static and dynamic analysis.

Numerous methods of crossing erection are available in practice and the decision for the solution is taken upon the security and economic aspects. The forces in the cable elements and bending moment in the pipeline are dependent on the steps of structure assembling and a construction stage analysis may reveal critical stages of force development.

Results of numerical simulations for a suspension crossing with the span of 160m considering static and dynamic actions are presented within the contents of this paper. The numerical model was defined by means of members: cables were modeled as cable elements taking into account initial stress, 2<sup>nd</sup> order geometrical nonlinearity and beam local nonlinearity, towers' elements and the pipe was modeled using linear bar elements.

The paper will give the results of the analysis of the same structure taking into consideration different masses for modal analysis. A real wind velocity-time variation as recorded on site is used for establishing the dynamic response of the structure to wind action. Also a simulation on the staged construction for the suspension crossing is performed.

This publication was supported by the European social fund within the framework of realizing the project „**Support of inter-sectoral mobility and quality enhancement of research teams at Czech Technical University in Prague**“, CZ.1.07/2.3.00/30.0034.

I. Both is with the Czech Technical University in Prague, Prague, 16629 Czech Republic (corresponding author, phone: 0040-727882621; e-mail: ioan.both@ct.upt.ro).

A. Ivan is with the Politehnica University of Timisoara, Timisoara, Timisoara, 300006 Romania (e-mail: adrian.ivan@upt.ro).

II. CASE STUDY

The structure analyzed in this paper represents a crossing with a span of 160m and two adjacent spans of 35m, an initial deflection of suspension cable of 12.5m leading to a value of 12.8 for the span to cable sag ratio. The vertical hangers are positioned at each 5m and the pipeline has a circular hollow section with a diameter of 700mm and a thickness of 8mm.

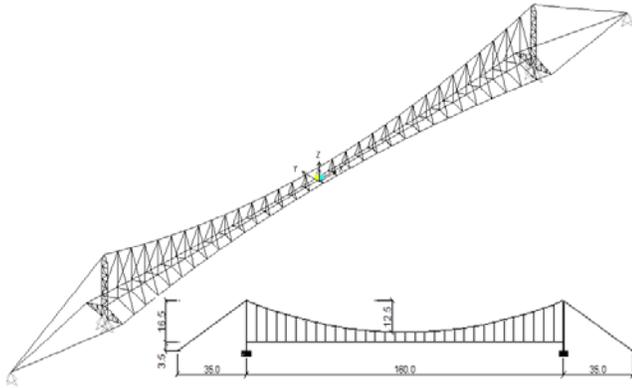


Fig. 1 Model of the case study

There are two main cables in the vertical planes situated at 2m and two inclined cables for lateral load and vertical stabilization. Two towers are placed at the end of the pipeline consisting of hot-rolled profiles HEB400 as vertical elements and square hollow sections 150/4 as bracings. The hangers and the wind ties have a diameter of 40mm whereas the main cables and the lateral cables have a diameter of 60mm (Fig. 1).

Such structures are highly complex since flexible support system may be attributed to the structure and for each element the boundary conditions are determined and influenced by the characteristics of the linked element [4]. Due to length variation each node of the main cable will have different interactions for the boundary conditions.

The cross section and material properties of the crossing are presented in Table I, where: M-main, L-lateral, T-torsional moment of inertia, E-modulus of elasticity, A-area of the cross-section,  $I_x$ -moment of inertia with respect to x axis,  $I_y$ -moment of inertia with respect to y axis. The properties of the tower are given for the entire truss structure.

Table I

| Element | E<br>[N/mm <sup>2</sup> ] | A<br>[mm <sup>2</sup> ] | T<br>[mm <sup>4</sup> ] | $I_x$<br>[mm <sup>4</sup> ] | $I_y$<br>[mm <sup>4</sup> ] |
|---------|---------------------------|-------------------------|-------------------------|-----------------------------|-----------------------------|
| M cable | 1.5-2e5                   | 2827                    | 0                       | 0                           | 0                           |
| L cable | 1.5-2e5                   | 2827                    | 0                       | 0                           | 0                           |
| Hangers | 1.5-2e5                   | 1194                    | 0                       | 0                           | 0                           |
| Windguy | 1.5-2e5                   | 1194                    | 0                       | 0                           | 0                           |
| Tower   | 2.1e5                     | 46.6e3                  | 3.17e5                  | 4.7e10                      | 1.2e9                       |
| Pipe    | 2.1e5                     | 17.4e3                  | 2.8e9                   | 1.04e9                      | 1.04e9                      |

In the analysis of the structure the following values of loading were considered:

- self-weight – program computed,
- permanent load - 0.7kN/m,
- imposed load from fluid - 3.675kN/m,

- wind - 0.46kN/m,
- pretension - variable.

All supports were considered to follow the restrictions only for translational degrees of freedom and the elements of the



Fig. 2 Wind recording station VS425

tower were also considered to be hinges.

For a dynamic analysis a load case defined function of the recordings of the wind action was considered. The values of

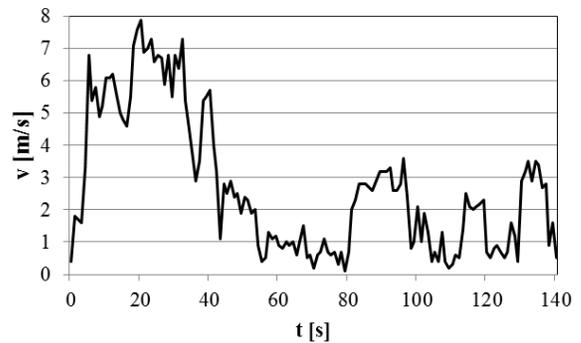


Fig. 3 Wind velocity

the wind velocities were provided by the local weather station for the west region of Romania monitored by an automatic station with ultrasonic transducer VS425 (Fig. 2) on the Mures river in Arad county, Varadia area. The maximum values recorded since the weather station was installed, 2009, are depicted in Fig.3.

III. LIMIT STATE ANALYSIS OF STRUCTURE

The forces and moments in the structural elements were determined using the FEM computer code SAP2000 for both static and dynamic analysis.

Damping ratios of such tension bar systems are particular due to their range interval as shown in [5]. It is common to use for a dynamic analysis proportional damping values from 0.04 to 0.1.

A certain level of pretension force is introduced in the cables of the structure in order to obtain the desired geometry in the final stage of construction. The Eigen modes of the

Due to the difference of 1.6s between the two cases of considered structural mass the resulting forces may vary

significantly. The behavior of structure taking into consideration the equivalent mass of self-weight is shown in Fig. 4 and Fig. 5

Fig. 4 shows the displacement of pipeline nodes in the horizontal plane for the dynamic analysis of wind action, applying the modal matrix for the pretensioned state of structure.

Fig. 5 shows the displacement of pipeline nodes in the horizontal plane for the dynamic analysis of wind action,

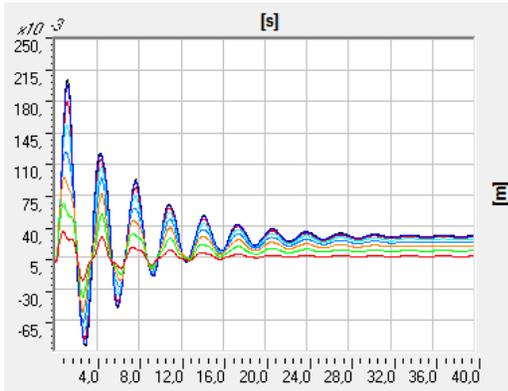


Fig. 4 Deflection - pretensioned state of structure ( $m_{perm}$ )

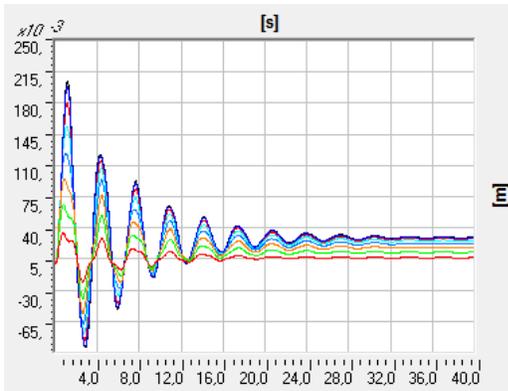


Fig. 5 Deflection - loaded state of structure ( $m_{perm}$ )

applying the modal matrix for the pretensioned and filled pipe, state of structure.

The curve following the maximum values in the chart of Fig. 4 – Fig. 7 represent the midspan of the crossing whereas the intermediate curves represent nodes between midspan and towers.

Table II

| Mode | $m_{perm}$                 |                         | $m_{perm+cvp}$             |                         |
|------|----------------------------|-------------------------|----------------------------|-------------------------|
|      | T [s]<br>( $K_{pretens}$ ) | T [s]<br>( $K_{fill}$ ) | T [s]<br>( $K_{pretens}$ ) | T [s]<br>( $K_{fill}$ ) |
| 1    | 3.2568                     | 3.0821                  | 4.8562                     | 4.6177                  |
| 2    | 2.1806                     | 2.2154                  | 3.2711                     | 3.3525                  |
| 3    | 2.0228                     | 1.8336                  | 2.9839                     | 2.7400                  |
| 4    | 1.8965                     | 1.7298                  | 2.8045                     | 2.5675                  |
| 5    | 1.4258                     | 1.3731                  | 2.0976                     | 2.0711                  |
| 6    | 1.3193                     | 1.3669                  | 1.9664                     | 2.0195                  |

The behaviour of structure taking into consideration the equivalent mass of self-weight and imposed load is shown in Fig. 6 and Fig. 7.

Fig. 6 shows the displacement of pipeline nodes in the horizontal plane for the dynamic analysis of wind action,

applying the modal matrix for the pretension state of structure.

Fig. 7 shows the displacement of pipeline nodes in the horizontal plane for the dynamic analysis of wind action, applying the modal matrix for the pretension and filled pipe, state of structure.

It can be seen that the maximum deflection of central node, when the mass is taken only the self-weight, is larger than the

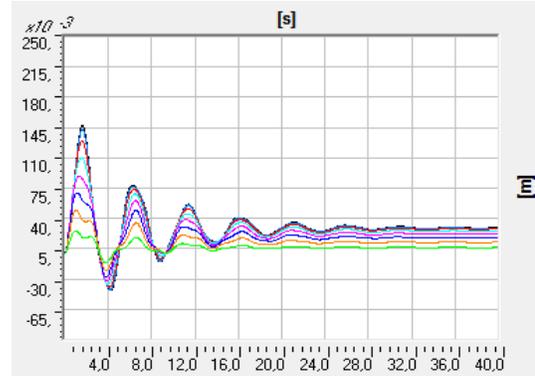


Fig. 6 Deflection - pretensioned state of structure ( $m_{perm+cvp}$ )

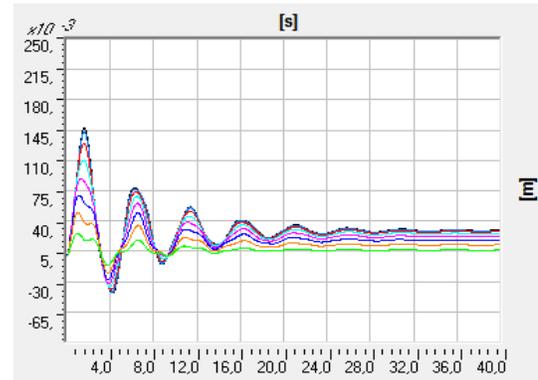


Fig. 7 Deflection - loaded state of structure ( $m_{perm+cvp}$ )

deflection in the case of equivalent mass from loaded pipeline, 200mm (Fig. 4) and 145mm (Fig. 6).

Not every node of the pipeline is following the sinusoidal path but their period (frequency) is the same for each node, therefore the first Eigen mode is relevant for analysis.

Table III Forces in the elements of the structure

| Comb. | Analysis | Main cable | Wind cable | Hanger | Anchor | Pipeline    |
|-------|----------|------------|------------|--------|--------|-------------|
|       |          | N [kN]     | N [kN]     | N [kN] | N [kN] | $M_z$ [kNm] |
| P+V   | Static   | 810.1      | 551.7      | 12.4   | 898.4  | 60.5        |
|       | Dynamic  | 805.8      | 444.2      | 12.6   | 893.8  | 21.6        |
| P+    | Static   | 1596.8     | 376.6      | 30.5   | 1759.3 | 63.2        |
|       | Dynamic  | 1593.0     | 390.9      | 30.6   | 1750.7 | 19.4        |

After 30s the deformed shape of the structure is stabilized and by comparing the figures where the mass is taken from the self-weight and the figures where the mass includes the loading (filling) of the pipeline, we can observe that the final displacement of the central node has the same value.

The remaining displacement is the result of the wind action that does not cause a dynamic effect on the structure.

All the above show how much the state of stress, used for dynamic analysis, influence the analysis results. Forces in cables have insignificant variation whereas the dynamic analysis leads to smaller bending moments in the pipeline

cross-section.

#### IV. CONSTRUCTION STAGE ANALYSIS

The erection of suspension structures is performed in stages where force values have a wide range of variation due to high deformations permitted by cables. As an effect of structural instability the structural system can change significantly and may lead to critical situations. Therefore a construction stage analysis is recommended to check the stability and stresses that may develop during intermediate stages.

For tension bar systems the effects of pre-tensioning forces are very important for the geometry during and after erection. According to design codes there has to be no compression forces in the cable elements of the structure. This would be the case when the finite elements used for modeling the cables are beam type elements.

In practice the errors caused by cable forces or mechanical properties are necessary to be monitored and corrected during construction stages. The main objectives of construction stage analysis may be summarized as follows:

- stress evaluation in cables for several stages,

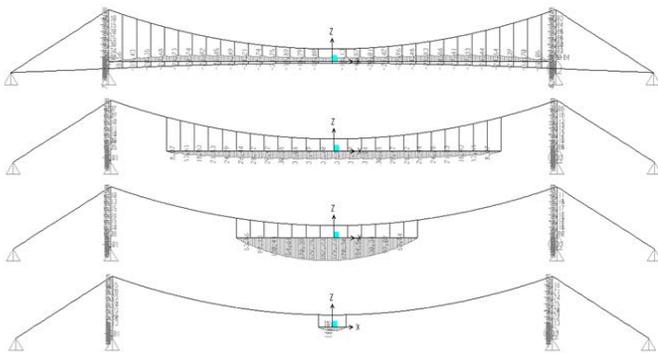


Fig. 8 Construction stages: midspan-towers

- geometric shape plotting related to final stage,
- leveling of pipeline,
- plastic deformations monitoring of elements during erection,
- stress and stability control in structural elements

Computer code SAP2000 allows the analysis of a structure in different stages of erection with the help of *Construction Scheduler* module. Function of the chronological order two possibilities of construction stage analysis may be defined: progressive or regressive. Resulting forces are very similar for each type of analysis but due to the already defined structure a regressive analysis is more convenient.

After defining the structure and the construction stage load case element groups of hangers and pipe elements must be defined according to the solution of assembling of structure. For the regressive analysis the entire structure has to be defined in a group and added in a construction stage. The following steps represent removal of groups of elements. Load action is allowed to be defined also with safety factors. Construction stage analysis may be defined with consideration of geometric non-linearity effects

A simple construction stage analysis may be defined with 5 phases [6] by defining groups of main parts of structural elements although it may be omitted critical situations as presented in the following.

There are multiple solutions for erecting the crossing to its final structural configuration and the forces are distributed accordingly. Fig. 8 presents the case of starting assembling the pipeline from the midspan of the crossing. Another possibility is to start the pipeline assembling from both ends as presented in Fig. 9 or just from one end, Fig. 10.

By performing a staged construction analysis for each of these cases it may be observed that for an intermediate phase the pipe is subjected to higher values of bending moments than as considered for the final stage. This situation is caused by the large deflections allowed by the main cable between

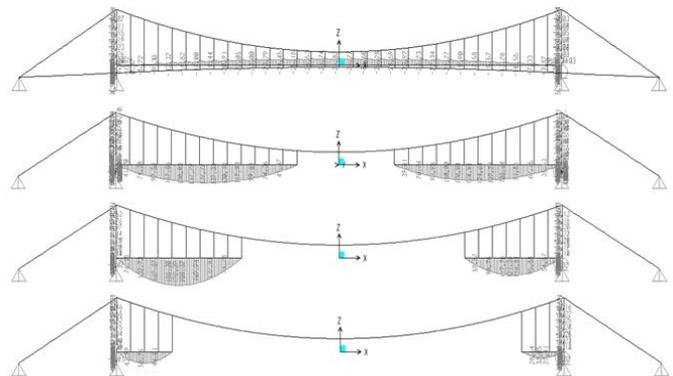


Fig. 9 Construction stages: towers - midspan

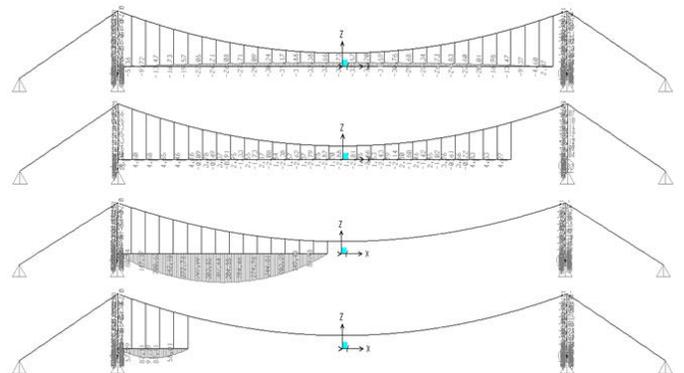


Fig. 10 Construction stages: tower - tower

the end connection points.

The maximum value for assembling the structure from the midspan is 182kNm and for the following two situations the circular section is subjected to 192kNm. These values represent almost double of the bearing capacity of the pipe cross-section.

As mentioned before the flexibility of the intermediate supports of the pipe allows high displacements leading to large equivalent spans. The high bending moments will result only if the pipeline is considered to be continuous. If the connections between adjacent pipe elements are not fully fixed the hinges resulted will reduce the bending moments. In a construction stage analysis the connection between elements cannot be modified therefore the results of an analysis that

includes simple connections between elements may only be simulated if groups of elements are not considered until the final stage of construction.

#### V. CONCLUSION

Analysis of structural systems with cables as tension elements exhibit a particular complexity owned by the large deformations resulted from the high flexibility of cables. The pretension of cables influences the Eigen-modes of the structure but a higher importance over the response to dynamic action is the consideration of the structure rigidity.

The unloaded structure and the case of filled pipeline define two situations that lead to rigidities of structure that are considered as an initial condition in dynamic analysis and lead to different values of deformations.

For a dynamic and static analysis the resulted forces in cable elements are similar whereas distinct values result for the pipe bending moment.

The construction stage analysis reveals that plastic deformation of pipeline may occur during intermediate phase of an erection method.

#### ACKNOWLEDGMENT

This publication was supported by the European social fund within the framework of realizing the project „**Support of inter-sectoral mobility and quality enhancement of research teams at Czech Technical University in Prague**“, CZ.1.07/2.3.00/30.0034.

#### REFERENCES

- [1] ASCE Manuals and Reports on Engineering Practice, Pipeline Crossings, No.89, 1996
- [2] Zhang Xin-jun, Sun Bing-nan, “Aerodynamic stability of cable-stayed-suspension hybrid bridges,” in *Journal of Zhejiang University Science*, No. 6A(8), 2005, p. 869-874
- [3] I.Both, “Dynamic response analysis of a pipeline crossing”, Bulletin of the Transilvania University of Braşov, pp.33-38, 2012
- [4] P. Cosmulescu, “Spatial Steel Structures”, Junimea, 1991.
- [5] Dragulinescu. M., Ghenoiu C., “Suspension crossings for pipelines, Revista Constructiilor si a materialelor de Constructii, no.9, vol. 14, 1962, p.447-456.
- [6] I. Both, A.Ivan, “Analiza pe faze de construcție a unei traversări suspendate”, a 12-a Conferința Națională de Construcții Metalice, 26-27 noiembrie 2010, Timisoara, Romania, pg. 129-134, ISBN 978-973-638-464-6, 2010

# Chromatics aberrations of diffractive elements in pulsed laser beams formation

Alexey P. Porfiriev, Sergey A. Degtyarev, Svetlana N. Khonina, Nikolai L. Kazanskiy

**Abstract**— The goal of our research is to create high-quality pulsed laser beams. We use diffractive optical elements (laser mode formers, diffractive gratings) for laser pulse beams formation. Diffractive optical elements have strong chromatic aberrations whereas short pulses have wide frequency spectrum. It is shown that low-indexed mode formers are resistant to chromatic dispersion. Experiments with tunable laser show the ability of using TEM(1,0) and TEM(1,1) mode formers for focusing into a set of closely-spaced light spots.

**Keywords**— chromatics aberration, diffractive element, mode former, pulse laser, tunable laser.

## I. INTRODUCTION

One of the greatest achievements of laser technology is the generation of very short laser pulses, which find wide expansion in material processing, microstructuring, laser filamentation, control of charged particles flow [1-5].

A lot of control methods of pulse time dependence have been created till this time. But in applications the spatial structure of a pulse is also important. Classical elements (lenses and mirrors) [6,7] offer quiet modest facilities for modulation of beam spatial structure. Interference schemes [8-10] are also used for generation fixed set of periodical spatial beam structures.

Most of spatial transformations of a laser beam are provided by diffractive optics equipment [11, 12]. Diffractive gratings and spatial light modulator (SLM) are used in most cases [13-18]. SLM can work dynamically but it has worse diffractive efficiency and spatial resolution.

However the structure of diffractive elements is optimal for only one fixed wavelength and it works properly only for a monochromatic beam. Spectral dispersion of a short laser pulse causes losses of image quality.

Different methods are offered for avoiding this effect. Particularly [13, 19], authors use two elements. In [13] it is diffractive and refractive one, and in [19] they use two diffractive elements: focusing one and diverging one. Also in

[20, 21] they use more complex chromatic aberration compensator.

At the same time in [22] authors show that phase diffractive elements are resistant to small departure of incident wavelength from planned calculated wavelength in the case of the elements correspond to laser modes.

In this work we provide experiments with different mode formers to study their resistance to chromatics dispersion. We take into consideration images which are created by the DOEs while it is illuminated by tunable laser.

## II. EXPERIMENTAL STUDY OF CHROMATICS DISPERSION INFLUENCE TO FOCAL IMAGE WHICH IS ACHIEVED WITH LOW-CODED DOE

To experimentally form Hermit-Gaussian mode TEM(1,1) we use phase doe with transmission function, which is showed in fig. 1. Very low coding level is used only to avoid rubbish out-aperture radiation.

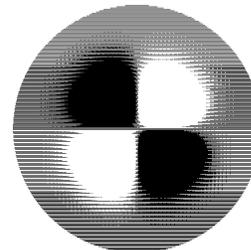


Fig. 1. Phase transmission function of low-coded diffractive element, which forms Hermit-Gaussian mod TEM(1,1)

A setup of conducted experiments is shown in fig. 2.

We use tunable laser with diode pumping NT200 which is made by EKSPLA corp. for generation a beam with a certain wavelength. Light filter F is used for attenuation of incident light. The set of microobjective MO1 (40x, NA = 0.6), lens L1 (f1=150 mm) and pinhole PH (hole size is 40 μm) serves to high-frequency filtration because laser produces a beam with bad quality. So after this set we achieve a quiet good Gaussian beam.

This work was supported by the Ministry of Education and Science of the Russian Federation.

A.P. Porfiriev (lporfiriev@rambler.ru) S.A. Degtyarev (sealek@gmail.com) are the Samara State Aerospace University, 34, Moskovskoye shosse, Samara, 443086, Russia.

S.N. Khonina and N.L. Kazanskiy are with Image Processing System Institute, 151, Molodogvardeyskaya street, Samara, 443001, Russia;

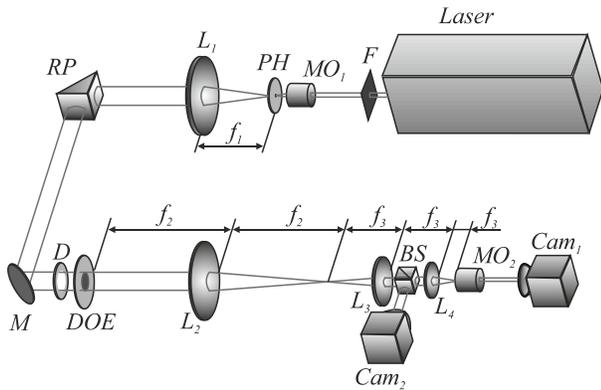


Fig. 2. Experimental optical scheme: Laser – laser with tunable wavelength EKSPLA NT200, F – light filter MO<sub>1</sub> – microobjective (40x, NA=0.6), PH – pinhole (40 mkm), L<sub>1</sub>, L<sub>2</sub>, L<sub>3</sub>, L<sub>4</sub> – lenses with focal lengths of  $f_1=150$  mm,  $f_2=250$  mm,  $f_3=100$  mm,  $f_4=60$  mm, D – iris diaphragm, DOE – diffractive optical element, BS – beam splitter, RP – rectangular prism, M – mirror, MO<sub>2</sub> – microobjective (40x, NA=0.6), Cam<sub>1</sub>, Cam<sub>2</sub> – CMOS-cameras MDCE-5A (1280x1024)

Beside that it is used for collimation of the beam to expand it to the element size. Collimated beam is directed to DOE with rectangular prism RP and mirror M. The pair of lenses L<sub>2</sub> ( $f_2=350$  mm) and L<sub>3</sub> ( $f_3=100$  mm) serves as telescopic system for imaging the DOE conjugate plane in plane of the lens L<sub>4</sub> ( $f_4=60$  mm). CMOS-cameras MDCE-5A (1/2", resolution 1280x1024 pixels) Cam<sub>1</sub> and Cam<sub>2</sub> are used for recording intensity distribution in conjugated DOE plane and in the focus of lens L<sub>4</sub>.

We conduct series of experiments to study the influence of wavelength upon the image in the focus of lens L<sub>4</sub>. For doing that we vary wavelength of tunable laser and investigate the image in the focus of lens L<sub>4</sub>. Wavelength can be set by control panel of tunable laser and image is received by camera Cam<sub>1</sub>. DOE diameter is 15 mm. Diameter of incident beam is equal to DOE diameter. Experimental results are shown in fig. 3. We can see that distributions for different wavelengths are equal in a qualitative sense. But the images are formed at different distances from the element for different wavelengths.

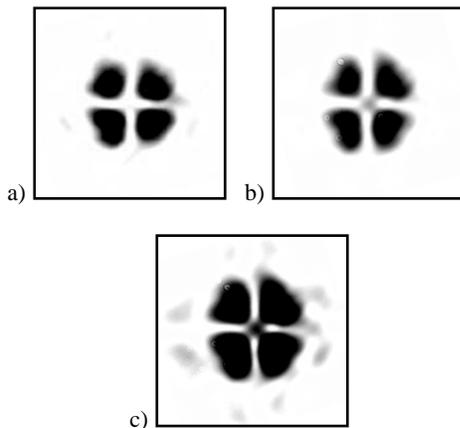


Fig. 3. Image of naturally formed intensity distributions of TEM<sub>11</sub> (negatives) in focal plane of lens L<sub>4</sub> for different wavelengths a) 846 nm, b) 906 nm; c) 946 nm

In the third series of experiments we research the influence of nonaxiality of the incident beam and DOE to the image in

the focal plane of lens L<sub>4</sub>. Intensity distributions for different types of improper illumination are shown in fig. 4 and 5. For varying of beam radius we use iris diaphragm D. In this experiments we take photos of both intensity distributions in focal plane of lens L<sub>4</sub> (using Cam<sub>1</sub>) and in the conjugated DOE plane (using Cam<sub>2</sub>). Results of the experiments are shown in fig. 4 to study the influence of inconsistency between diameter of incident beam and diameter of DOE for different wavelength. We achieve proper Hermit-Gaussian mode TEM(1,1) in case of DOE is fully illuminated (sizes of the beam and the element are equal) and when the incident beam and the element are coaxial. But in case of the beam size is greater than the element, the image distorts. It happens because a part of beam passes beside the element and then it is focused by the lens L<sub>4</sub> into the image. To decrease this effect we use low-level phase coding.

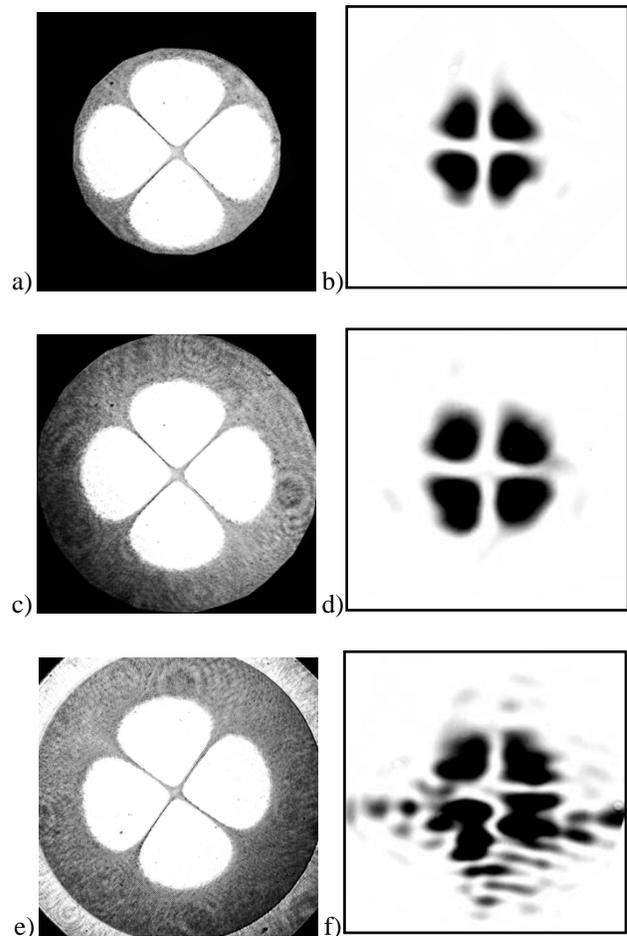


Fig. 4. Image of a plane, that is conjugated the DOE plane. DOE serves for Hermit-Gaussian mode TEM(1,1) forming (left column) and experimentally formed intensity distributions (right column, negative) in the focal plane of the lens L<sub>4</sub>. Different sizes of the incident beam are varied by changing of iris-diaphragm radius. Wavelength is of 846 nm.

In fig. 5 incident beam is shifted to upper-left corner (a) so image (b) is distorted. An intensity maximum also appears in the center of picture. The artefact must not arise among the four spots of proper mode TEM(1,1).

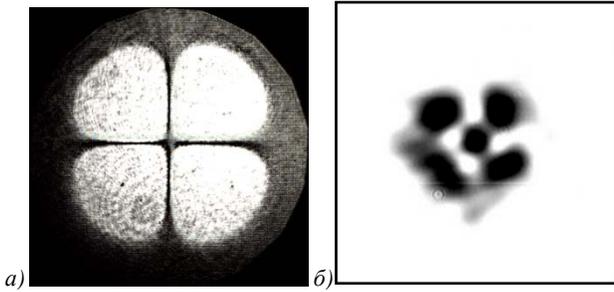


Fig. 5. Work of DOE for forming of Hermit-Gaussian mode TEM(1,1) when incident beam and the element are nonaxial. Picture of conjugated DOE plane (incident beam is shifted to upper left corner) (a); experimentally formed picture of intensity distribution in focal plane of lens  $L_4$  (b); wavelength is of 846 nm

### III. DYNAMICAL CONTROL OF GAUSSIAN BEAM FOCUSING BY INCIDENT WAVELENGTH

We have studied one interesting moment, that DOE chromatism makes it possible to manage the focal picture by means of varying incident wavelength. This facts follows from mismatch between incident wavelength and relief height of DOE [22]. But it is obvious, that the shape of the beam changes not only in the focal plane, but also in another planes at different distances from the DOE.

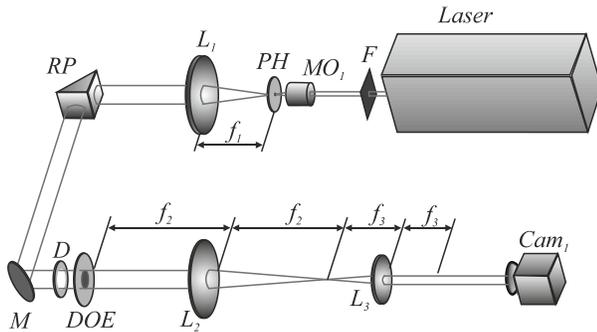


Fig. 6. Experimental optical setup. Laser – tunable laser EKSPLA NT200, F – light filter MO1 – microobjective (40x, NA=0.6), PH– pinhole (40 mkm), L1, L2, L3– lenses with focal distances are of  $f_1=150$  mm,  $f_2=250$  mm,  $f_3=100$  mm, D – iris diaphragm, DOE – diffractive optical element, BS – beam splitter, RP – rectangular prism, M – mirror, Cam1 – CMOS-camera MDCE-5A (resolution is 1280x1024)

To demonstrate dynamical beam reshaping we use experimental setup, which is shown in fig. 6. Used 8-sectors DOE is shown in fig. 7. Phase jump between adjacent sectors is  $\pi$  radian when wavelength is of 800 nm. We vary the wavelength in increments of 50 nm and observe an intensity distribution at a distance of 300 mm from the lens  $L_3$ . Experimentally received pictures are shown in fig. 8. We can see, that if wavelength is of 800 nm, there are 8 spots in the observed plane. But the shape of the beam is changing while the wavelength is varying from 600 nm to 1000 nm. Four peaks are becoming weaker and four peaks are becoming stronger. When the wavelength is 1000 nm there are 4 peaks in the observed plane instead of 8 peaks (fig. 8 e).

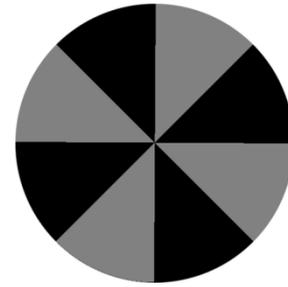


Fig. 7. Phase function of the 8-sectors diffractive optical element. Gray – 0 radian, Black –  $\pi$  radian.

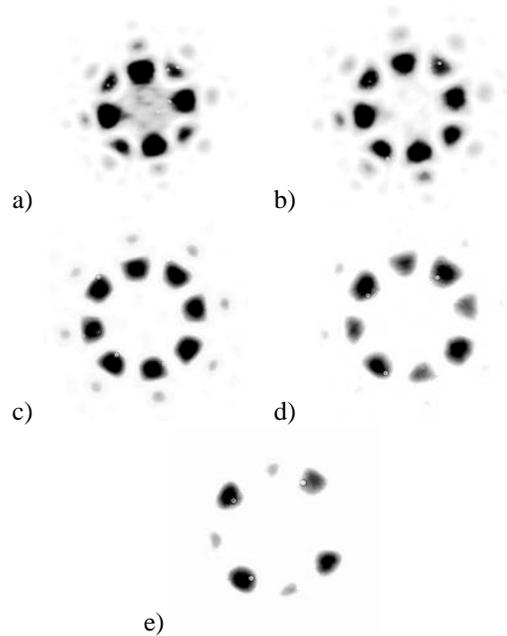


Fig. 7. Phase function of the 8-sectors diffractive optical element. Gray – 0 radian, Black –  $\pi$  radian.

### IV. CONCLUSION

Experimental study with tunable laser has shown that diffractive TEM(1,0) and TEM(1,1) modes formers can be used for beam focusing into closely-spaced light spots. High quality of the image and quiet chromatics dispersion resistance are provided for quiet wide wavelength spectrum. For higher-order mode additional coding is required. But in this case diffractive efficiency is falling down, chromatics distortions are growing up and a lot of artefacts are arising.

In addition we can dynamically manage the intensity picture through varying the wavelength of tunable laser. In this case we use only one mode former which corresponds to the certain wavelength. But for another wavelength received picture is different. We show that 8-sectors DOE produces 8 peaks for 800 nm wavelength of incident beam. But for 1000 nm in produces 4 peaks.

### ACKNOWLEDGMENT

This work was funded by the Ministry of Education and Science of the Russian Federation.

## REFERENCES

- [1] Umstadter, D., "Relativistic laser-plasma interactions," *Journal of Physics D: Applied Physics* 36(8), 151–165 (2003).
- [2] Sun, H.-B., Kawata, S., "Two-photon photopolymerization and 3D lithographic microfabrication," *Advances in Polymer Science* 170, 169–273 (2004).
- [3] Salamin, Y.I., Hu, S.X., Hatsagortsyan, K.Z., Keitel, C.H., "Relativistic high-power laser-matter interactions," *Physics Reports* 427(2-3), 41 – 155 (2006).
- [4] Malka, V., Faure, J., Gauduel, Y.A., Lefebvre, E., Rousse, A., Phuoc, K.T., "Principles and applications of compact laser-plasma accelerators," *Nature Physics* 4, 447–453 (2008).
- [5] Cheng, J., Liu, C., Shang, S., Liu, D., Perrie, W., Dearden, G., Watkins, K., "A review of ultrafast laser materials micromachining," *Optics and Laser Technology* 46, 88–102 (2013).
- [6] Kato, N., Takeyasu, N., Adachi, Y., Sun, H.-B., Kawata, S., "Multiple-spot parallel processing for laser micronanofabrication," *Applied Physics Letters* 86(4), 044102–044104 (2005).
- [7] Salter, P.S., Booth, M. J., "Addressable microlens array for parallel laser microfabrication," *Optics Letters* 36(12), 2302–2304 (2011).
- [8] Shoji, S., Kawata, S., "Photofabrication of three-dimensional photonic crystals by multibeam laser interference into a photopolymerizable resin," *Applied Physics Letters* 76(19), 2668–2670 (2000).
- [9] Kondo, T., Matsuo, S., Juodkakis, S., Mizeikis, V., Misawa, H., "Multiphoton fabrication of periodic structures by multibeam interference of femtosecond pulses," *Applied Physics Letters* 82(17), 2758–2760 (2003).
- [10] Dong, X.-Z., Zhao, Z.-S., Duan, X.-M., "Micronanofabrication of assembled three-dimensional microstructures by designable multiple beams multiphoton processing," *Applied Physics Letters* 91(12), 124103 (2007).
- [11] Soifer, V. A., [Computer Design of Diffractive Optics], Cambridge Inter. Scien. Pub. Ltd. & Woodhead Pub. Ltd. (2012).
- [12] Soifer, V. A., [Diffractive Nanophotonics], Cambridge Inter. Scien. Pub. Ltd. & Woodhead Pub. Ltd. (2014).
- [13] Kuroiwa, Y., Takeshima, N., Narita, Y., Tanaka, S., Hirao, K., "Arbitrary micropatterning method in femtosecond laser microprocessing using diffractive optical elements," *Optics Express* 12(9), 1908–1915 (2004).
- [14] Hayasaki, Y., Sugimoto, T., Takita, A., Nishida, N., "Variable holographic femtosecond laser processing by use of a spatial light modulator," *Applied Physics Letters* 87(3), 031101 (2005).
- [15] Kelemen, L., Valkai, S., Ormos, P., "Parallel photopolymerisation with complex light patterns generated by diffractive optical elements," *Optics Express* 15(22) 14488–14497 (2007).
- [16] Kuang, Z., Perrie, W., Leach, J., Sharp, M., Edwardson, S. P., Padgett, M., Dearden, G., Watkins, K. G., "High throughput diffractive multi-beam femtosecond laser processing using a spatial light modulator," *Applied Surface Science* 255(5), 2284–2289 (2008).
- [17] Kuang, Z., Liu, D., Perrie, W., Edwardson, S., Sharp, M., Fearon, E., Dearden, G., Watkins, K., "Fast parallel diffractive multi-beam femtosecond laser surface micro-structuring," *Applied Surface Science* 255(13-14), 6582–9588 (2009).
- [18] Obata, K., Koch, J., Hinze, U., Chichkov, B. N., "Multi-focus two-photon polymerization technique based on individually controlled phase modulation," *Optics Express* 18(16), 17193–17200 (2010).
- [19] Amako, J., Nagasaka, K., Kazuhiro, N., "Chromatic-distortion compensation in splitting and focusing of femtosecond pulses by use of a pair of diffractive optical elements," *Optics Letters* 27(11), 969–971, (2002).
- [20] Torres-Peiró, S., González-Ausejo, J., Mendoza-Yero, O., Mínguez-Vega, G., Andrés, P., Lancis, J., "Parallel laser micromachining based on diffractive optical elements with dispersion compensated femtosecond pulses," *Optics Express* 21(26), 31830–31836 (2013).
- [21] Zapata-Rodríguez, C.J., Caballero, M.T., "Isotropic compensation of diffraction-driven angular dispersion," *Optics Letters* 32(17), 2472–2474, (2007).
- [22] Karpeev, S.V., Alferov, S.V., Khonina, S.N., Kudryashov, S.I., "Study of the broadband radiation intensity distribution formed by diffractive optical elements," *Computer Optics* 38(4), 689–694 (2014).

# Polarization angle independent perfect multi-band metamaterial absorber in microwave frequency regime

O. T. Gunduz and C. Sabah

**Abstract**—A new kind of multi-band metamaterial absorber is engineered to be used in novel microwave components and devices. The structure is designed based on the concentric ring resonator in which each ring provides a resonance at different frequencies. Numerical investigations are carried out step by step for being able to observe the effect of each ring resonator in the studied frequency range. The absorption character of the proposed structure is also investigated for the change of the polarization angle. Our results reveal that the structure almost perfectly absorbs the electromagnetic wave energy at multiple resonant frequencies in the microwave range. The accomplishment of polarization angle independency with maintaining the multi-band behavior provides a huge availability for the metamaterial absorber to be used in myriad applications. The number of resonances can be increased by adding more resonators with more splits concentrically and this will also guide the researchers to design more advanced multi-band absorbers.

**Keywords**— Metamaterial, absorber, microwave, multi-band.

## I. INTRODUCTION

IN 1967, Veselago proposed [1] that it is possible to create artificial materials whose relative permeability and permittivity are both negative simultaneously. However, to experiment actual materials was not possible at the time since these kind of properties had never been observed in the nature as Veselago cited in his 1967 paper. After 32 years, in their study, [2] Pendry and his colleagues analyzed some magnetic microstructures theoretically to show that how their magnetic properties could be improved by a split-ring resonator (SRR) structure, they proposed, responding as if they have an effective magnetic constant, whose real part was negative at particular frequencies, for microwave radiation and one year later, Smith and his colleagues published the first article to have the experimental investigation of an SRR [3] revealing peculiar properties about single and double negative (SNG and DNG) metamaterials in microwave region. Their structure was a “left-handed” material with which the phenomena of Doppler

Effect, Cherenkov Effect and Snell’s law were reversed as Veselago predicted. In the light of these studies, the power of left-handed materials on controlling electromagnetic (EM) waves has been proved and their propagation characteristic features have been mastered. After, those studies many resonator designs which were the analogic replicas of the SRR structure settled in the literature and gained enormous attention in these days. Also, open-ring resonators (ORRs) topology is become another most preferred type of resonator [4]. Depending on these two categories, different type of resonators in different geometrical shapes and unique resonance properties are enhanced at present. Owing to their capacitive and inductive effects, it is now possible to direct the energy density of the joint oscillation of electron density through the required points on the surface of the resonator, by just simply adjusting the dimensions, so as to obtain almost maximum energy in the form of EM radiation without both reflection and transmission to occur. Thus, this logical structure provides nearly maximum absorption at some resonant frequencies. Not only varying the size of the novel structure could lead a much better absorption level to appear but also increasing the number of SRRs or ORRs could increase the number of resonant frequencies at which useful absorption peaks occur [4].

In this paper, a new kind of multi-band metamaterial absorber (MA) is designed and enhanced so as to absorb electromagnetic energy in a particular region of the spectrum, especially in microwave region. The design includes three squared-shape resonators nested. Separately, each squared-shape resonator forms a mono band absorber with the substrate of the multi-band absorber and all together, they form the proposed multi-band structure responding nine different resonant frequencies at 5.96 GHz, 8.48 GHz, 11.02 GHz, 13.64 GHz, 14.98 GHz, 16.78 GHz, 20.28 GHz, 21.7 GHz, and 22.98 GHz with the four of these frequencies have almost perfect absorption of 99.77%, 95.59%, 99.75%, and 99.42% between the frequency spectrums of interest respectively. Our goal is to reach the multi-band case by analyzing each mono-band absorber formed by each resonator individually and to show how the effect of their specific combinations on the response of each resonator improves the absorption numbers in order to create the four main absorption peaks of the suggested multi-band MA absorber.

The work reported here was carried out at Middle East Technical University — Northern Cyprus Campus (METU-NCC). It is supported by METU-NCC under the grant number of BAP-FEN-13-D-4.

O. T. Gunduz is with the Middle East Technical University - Northern Cyprus Campus, Department of Electrical and Electronics Engineering, Kalkanli, Guzelyurt, 99738, TRNC / Mersin 10, Turkey.

C. Sabah is with the Middle East Technical University - Northern Cyprus Campus, Department of Electrical and Electronics Engineering, Kalkanli, Guzelyurt, 99738, TRNC / Mersin 10, Turkey (sabah@metu.edu.tr).

## II. DESIGN AND SIMULATION

In this study, we present a method to develop a multi-band MA step by step. The designed square-ring resonator (ASRR), has three different square-shaped resonators each corresponding three different resonance frequencies. It can be confidently claimed that the number of resonators determines the number of different resonance frequencies [4]. However, it is not only the number of resonators that determines the number of different resonance peaks in the magnitude spectrum of transmission but also dimensions, number and orientations of gaps of whole structure is playing a huge role in the engineering process of spectral positioning and bandwidth of the resonances [5].

Fig. 1 Shows the three dimensional geometry with square-shaped resonators (SSR) of the proposed multi-band absorber with its three dimensional orientation. The suggested device is engineered by using FR4 ( $\epsilon=4.3$  and  $\tan\delta=0.025$ ) as the substrate with sizes 18mm in the x-, y-, directions and with the thickness of 1mm in the z- axis. The metal made up the SSRs is copper the thickness of 0.036mm and the conductivity of  $5.8 \times 10^7$  S/m. The gaps between the resonators and line widths are 1mm. The size of the inner, middle, and outer resonators in x- and y- directions are 8mm, 12mm, and 16mm respectively. Each resonator has four identical spaces having widths of 1mm.

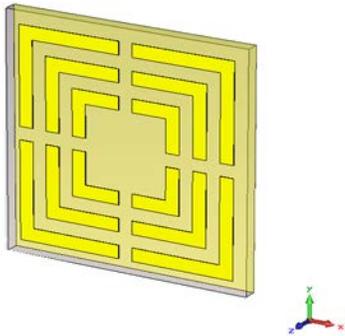


Fig. 1 Geometry of multiband metamaterial absorber.

Numerical simulations are held by using a commercial full-wave EM solver based on finite integration technique to obtain the frequency responses of reflectance  $R(\omega) = |S_{11}|^2$  so as to determine the absorption characteristics of the material [6]. There is no need to obtain the characteristics of the transmission  $T(\omega) = |S_{21}|^2$  since there is a copper ground plane preventing even a tiny bit of transmission to occur. Knowing that all the incident wave to ground plane is reflected back, the only parameter which is left here to consider for better understanding of the absorption response is nothing but reflectance. Therefore, a waveguide port is placed along the z- axis for both energizing the material and observing the frequency spectrum of the reflection formations. Boundary conditions are assigned as periodic along x- and y- directions and open (add space) along the z-direction.

The entire structure is analyzed by considering each resonator alone on the dielectric substrate, in order to investigate their effects separately as shown in Fig. 2. Each resonator forms four resonance peaks related to the number of splits it has [5]. Therefore, it is logical to inference that there will be twelve resonance locations in total. In the chosen frequency range the only SRR which exhibits four of its resonance is only the outer resonator as shown in Fig. 2. The inner ring which is the smallest resonator of the proposed absorber causes a remarkable main resonance at 13.66 GHz with just over 99.99% of absorption of the incident wave propagation in Fig. 2 a). Even only the presence of the inner ring is enough in order the left-handed material to exhibit strong absorption characteristics in high frequencies. Other three resonances one of which is not included in the current frequency range are not much useful compared to the main dip at 13.66 GHz. The middle resonator is also successfully resonates similar to, which the inner SRR does, with a 99.52% of absorption percentage at 8.42 GHz as illustrated in Fig. 2 b). Besides, its secondary resonance at 14.76 GHz has at least 99% of electromagnetic power absorption although, the middle resonator is designed to harvest energy at the frequencies lower than 14.76 GHz. Lastly, the resulting frequency spectrum of  $S_{11}$  curve, demonstrating a quadruple-band resonance characteristics as the other two resonators do, is corresponding to the outer SRR presented in Fig. 2 c) with around 99.16% and 99.95% of absorption peaks at 5.98 GHz and 11.12 GHz respectively. Even though it is engineered for the lowest resonance frequency operation, the outer SRR has its strongest resonance dip at 11.12 GHz instead of 5.98 GHz. After combining all the resonators together on the same dielectric substrate, the desired frequency spectrum of  $S_{11}$  parameter is screened in Fig. 2 d) with the locations and the strengths of the resonance peaks. The simulation results of the reflection coefficient shown in Fig. 3 exactly coincide with the expected characteristics indicated by Fig. 2 a) and b) except few undesired location alterations and distortions. One little problem occurring here is that when all the resonators are added together, we encounter with certain tradeoffs for instance, the other high frequency components of the resonances are less efficient and become distorted in the multiple resonator case. The strongest possibility is that the coupling between the SRRs playing a huge role in the multiple formations and their distortive structures. If this is the case then these interactions could be lowered by locating additional lumped elements between the important points, having high possibility to strengthen the coupling effects among the SRRs. By tuning the parameters of the lumped elements, the resonance characteristics could be enhanced easily without playing with the dimensions of the material.

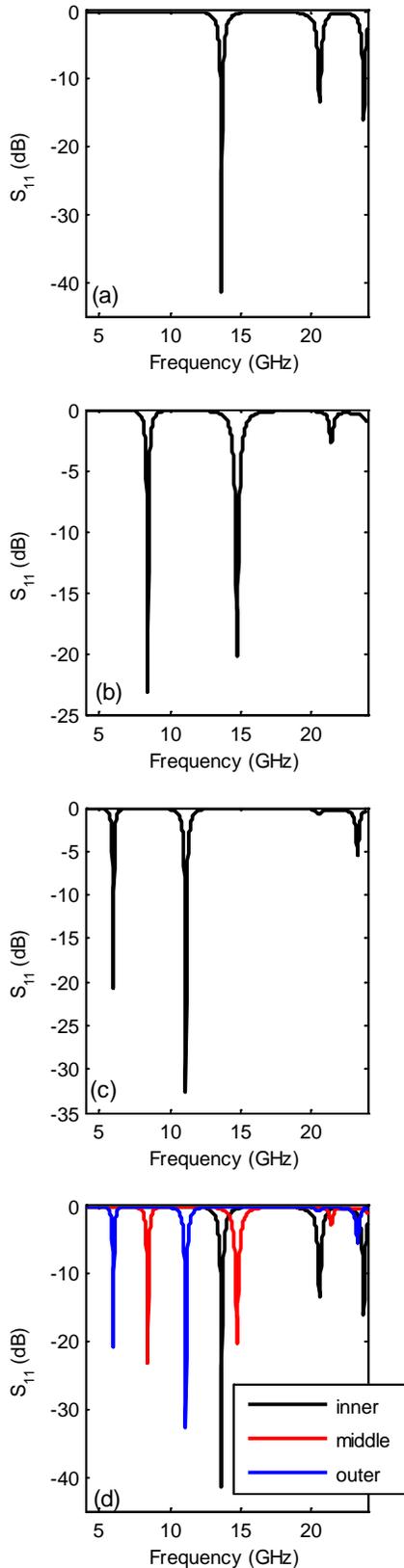


Fig. 2 a) The resonance frequencies of the inner SRR at 13.66 GHz, 20.6 GHz, and 23.72 GHz; b) The resonance frequencies corresponding to the middle SRR at 8.42 GHz, 14.76 GHz, and 21.42 GHz; c) The resonance frequencies corresponding to the outer SRR at 5.98 GHz, 11.12 GHz, 20.58 GHz, and 23.3 GHz; d) The expected frequency response of the combination of all resonators.

Despite these tradeoffs of the topology, the huge part of the absorption characteristics could be covered without a significant change in the resonance frequency locations and the amount of power absorbed by the left-handed material as it is confirmed by Fig. 3 a) and b).

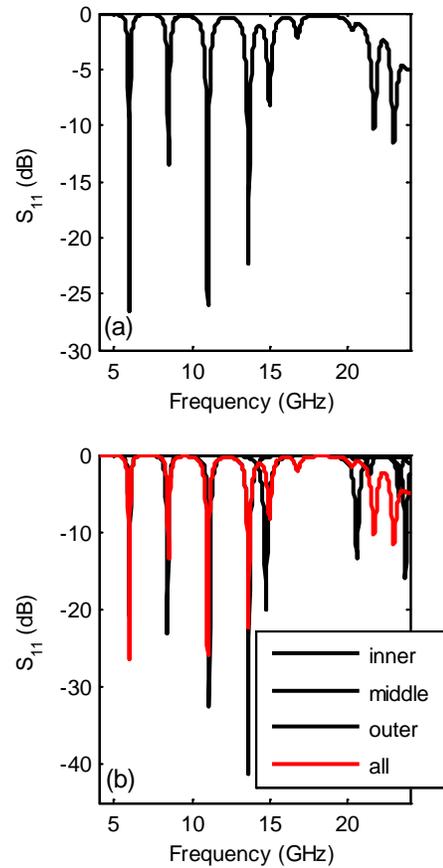


Fig. 3 a) The actual frequency response of the entire structure with the all SRRs b) the comparison of the expected and the actual multi-band absorber characteristics.

The proposed structure is also simulated through its polarization angle independency characteristics by varying the polarization angle  $\theta$  between 0 and 90 degrees. Fig. 4, reveals the success of the material to be polarization angle independent. Since the response is exactly same for all angle values as a result of a high resolution simulation with a huge number of hexahedral meshes, it is impossible to show each curve separately. They appear as if there is just a single curve in Fig. 4 indicating the same successful operation under various polarization angles [6-9].

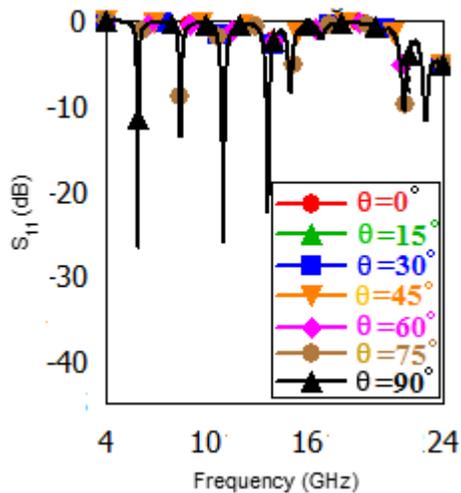


Fig. 4 The frequency response of  $S_{11}$  parameter for different linear polarization angle values (The rate of increment of the polarization angle is 15 degrees).

### III. CONCLUSION

A multi-band metamaterial absorber is structured and examined by means of single and multiple SRR combinations and polarization independency through simulations. Each SRR is treated separately and their combinational effects are also screened and compared with the desired dual-band frequency behavior of the absorber. According to the consequences of the numerical investigations, the suggested MA exhibits and maintains perfect absorption characteristics in both the single SRR and multi SRRs case for the certain frequency points in the spectrum. Also, it preserves the resonance frequency locations, which satisfies the purpose of combining SRRs together, when all the structure is gathered. In the case of polarization angle variation the proposed MA is stable and the resulting resonance frequencies remain roughly untouched in terms of the frequency and the absorption rate. The mentioned polarization angle independent MA is verified to be utilized as a multi-band absorber in microwave region of the electromagnetic spectrum and it can be useful in certain applications such as thermal detectors, stealth technology, spectroscopic imaging, sensors, etc.

### REFERENCES

- [1] V. G. Veselago, "The electrodynamics of substances with simultaneously negative values of  $\epsilon$  and  $\mu$ ," *Sov. Phys. Uspekhi*, vol. 10, no. 4, pp. 509-514, July 1968.
- [2] J. B. Pendry, A. J. Holden, D. J. Robbins, and W. J. Steward, "Magnetism from conductors and enhanced nonlinear phenomena," *IEEE T. Microw. Theory*, vol. 47, pp. 2075-2084, Nov. 1999.
- [3] D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, and S. Schultz, "Composite medium with simultaneously negative permeability and permittivity," *Phys. Rev. Lett.*, vol. 84, no. 18, pp. 4184-4187, May. 2000.
- [4] C. Sabah, "Multi-band metamaterials based on multiple concentric open-ring resonators topology," *IEEE J. Sel. Top. Quant.*, vol. 19, no. 1, pp. 8500808-1-8, Jan. 2013.

- [5] Q. Ye, Y. Liu, H. Lin, M. Li, H. Yang, "Multi-band metamaterial absorber made of multi-gap SRRs structure," *Applied Physics A*, vol. 107, no. 1, pp. 155-160, Feb. 2012.
- [6] N. I. Landy, S. Sajuyigbe, J. J. Mock, D. R. Smith, W. J. Padilla, "Perfect metamaterial absorber," *Phys. Rev. Lett.*, vol. 100, no. 20, pp. 207402-1-4, May. 2008.
- [7] F. Dincer, M. Karaaslan, E. Unal, K. Delihacioglu, and C. Sabah, "Design of polarization and incident angle insensitive dual-band metamaterial absorber based on isotropic resonator," *Prog. Electromagn. Res.*, vol. 144, pp. 123-132, Jan. 2014.
- [8] F. Dincer, O. Akgol, M. Karaaslan, E. Unal, and C. Sabah, "Polarization angle independent perfect metamaterial absorbers for solar cell applications in the microwave, infrared, and visible regime," *Prog. Electromagn. Res.*, vol. 144, pp. 93-101, Jan. 2014.
- [9] X. Shen, T. J. Cui, J. Zhao, H. F. Ma, W. X. Jiang, and H. Li, "Polarization-independent wide-angle triple-band metamaterial absorber," *Opt. Express*, vol. 19, no. 10, pp. 9401-9407, May. 2011.
- [10] L. Huang and H. Chen, "Multi-Band and Polarization Insensitive Metamaterial Absorber," *Prog. Electromagn. Res.*, vol. 113, pp. 103-110, Jan. 2011.

**O. T. Gunduz** is a student at the Middle East Technical University - Northern Cyprus Campus, Department of Electrical and Electronics Engineering. He is mainly working in the field of metamaterials and their applications.

**Cumali Sabah** received the B.Sc., M.Sc., and Ph.D. degrees in Electrical and Electronics engineering. He is currently with Middle East Technical University - Northern Cyprus Campus. His research interests include the microwave and electromagnetic investigation of unconventional materials and structures, wave propagation, scattering, complex media, metamaterials and their applications.

# The Simulation of Negative Influences in the Environment of Fixed Transmission Media

Rastislav Róka

**Abstract:** This lecture is devoted to the simulation of negative influences in the environment of fixed transmission media. There are two basic areas of fixed transmission environment – metallic and optical. An attention is focused on main features and characteristics of negative influences at the signal transmission. Consequently, simulation models for appropriate transmission paths are introduced with short description of functional blocks representing technologies utilized in the specific environment. The created Simulink modeling schemes of real environmental conditions at the signal transmission allow executing different requested analyses for advanced techniques of the digital signal processing.

**Keywords:** metallic homogeneous lines, power distribution cables, optical single-mode fibers, simulation models

## I. INTRODUCTION

FOR successful understanding of the signal transmission in access networks that utilized fixed transmission media, it is necessary exactly to recognize essential negative influences in the real environment of metallic homogeneous symmetric lines, power distribution cables and optical fibers. This lecture discusses features and frequency characteristics of negative influences on signals transmitted by means of the VDSL technology, the PLC technology and PON networks. For the expansion of communication systems on fixed transmission media, it is necessary to have a detailed knowledge of their transmission environments and negative influences in the real developing of customer installations.

A main attention of the metallic transmission environment is focused on the explanation of substantial negative influences and on the description of the proposed VDSL and PLC simulation models. Presented simulation models represent a reach enough knowledgebase for the extended digital signal processing techniques of the VDSL and PLC signal transmissions that can be extremely helpful for various tests and performance comparisons.

This work is a part of research activities conducted at Slovak University of Technology Bratislava, Faculty of Electrical Engineering and Information Technology, Institute of Telecommunications, within the scope of the project KEGA No. 039STU-4/2013 “Utilization of Web-based Training and Learning Systems at the Development of New Educational Programs in the Area of Optical Transmission Media”.

Rastislav Róka is with the Institute of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia (e-mail: rastislav.roka@stuba.sk).

A main attention of the optical transmission environment is focused on the explanation of its substantial linear and nonlinear effects and on the description of the proposed optical communication path’s simulation model. The presented simulation model represents a reach enough knowledgebase that can be helpful for various tests and performance comparisons of various novel modulation and encoding techniques suggested and intended to be used at signal transmissions in the transmission environment of optical fibers.

## II. THE ENVIRONMENT OF METALLIC HOMOGENEOUS LINES

### A. Linear negative influences on transmitted signals

Propagation loss and linear distortions (distortions of the module and the phase characteristics and the group delay characteristic) are linear negative influences dependent on physical and constructional parameters, such as a line length, a core diameter of the wire, a mismatch of impedances in cross-connecting points of sections, a frequency bandwidth and so forth [13].

We first discuss the propagation loss  $L_{dB}$  in a perfectly terminated line. If  $R$ ,  $L$ ,  $G$  and  $C$  are primary constants of the line and  $\omega = 2\pi f$ , where  $f$  is the frequency, then

$$\gamma(\omega) = \alpha(\omega) + j\beta(\omega) = \sqrt{(R + j\omega L)(G + j\omega C)} \quad (1)$$

and

$$\mathcal{Z}(\omega) = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \quad (2)$$

where  $\gamma(\omega)$  denotes the propagation constant,  $\alpha(\omega)$  is the specific attenuation constant,  $\beta(\omega)$  is the specific phase-shift constant and  $\mathcal{Z}(\omega)$  is the characteristic impedance. For a perfectly terminated homogeneous line with the length  $l$ , the transfer function  $\mathcal{H}(l, f)$  of metallic homogeneous symmetric lines is given by

$$\mathcal{H}(l, f) = e^{-l\gamma(f)} = e^{-l\alpha(f)} \cdot e^{-j l \beta(f)} \quad (3)$$

and the propagation loss  $L_{dB}$  is given as

$$L_{dB}(l, f) = -20 \cdot \log_{10} |\mathcal{H}(l, f)| = \frac{20}{\ln 10} \cdot l \cdot \alpha(f) = a_{line}(l, f) [dB] \quad (4)$$

We must place emphasis on the interchangeable use of terms - the line attenuation  $a_{line}(l, f)$  and the propagation loss  $L_{dB}(l, f)$  to designate the quantity in (4) only for the case of a perfectly terminated line. We can see that a dependency of the propagation loss  $L_{dB}$  on the line length  $l$  is linear and is also an increasing function of the frequency  $f$  as should be apparent from the expression for the propagation constant  $\gamma(\omega)$  in (1). A power level of transmitted signals is also influenced by other important parameters - a diameter and constructional material of the core.

### B. Near-end and Far-end crosstalk signals

The term ‘‘crosstalk’’ generally refers to the interference that enters a communication channel through some coupling paths. If either single or multiple interferers generate a crosstalk signal, we can define a gain of the NEXT crosstalk path using a following relation

$$|\mathcal{H}_{NEXT}(l, f)|^2 = \frac{\pi^2 \cdot f^2 \cdot k_{NEXT}}{\alpha(f)} [1 - e^{-4 \cdot \alpha(f) \cdot l}] \approx K_{NEXT} \cdot f^{3/2} \quad (5)$$

where variables are given as  $K_{NEXT} = 0,882 \cdot 10^{-14} \cdot N_d^{0,6}$ ,  $N_d$  is the number of disturbing pairs (disturbers),  $f$  is the frequency in Hz. An approximation on the right in (5) is valid when the line length  $l$  is large and for frequency regions where the real part  $\alpha(\omega)$  of the propagation constant is proportional to  $\sqrt{f}$ . We can also derive a gain of the FEXT crosstalk path in a similar manner using a following relation

$$|\mathcal{H}_{FEXT}(l, f)|^2 = 4 \cdot \pi^2 \cdot f^2 \cdot k_{FEXT} \cdot l \cdot e^{-2 \cdot \alpha(f) \cdot l} \approx K_{FEXT} \cdot l \cdot 3280 \cdot f^2 \cdot |\mathcal{H}(l, f)|^2 \quad (6)$$

where variables are given as  $K_{FEXT} = 3,083 \cdot 10^{-20}$ ,  $l$  is the line length in km,  $f$  is the frequency in Hz and  $\mathcal{H}(l, f)$  expresses the transfer function of a metallic homogeneous symmetric line.

From a data communication point of view, the NEXT crosstalk is generally more damaging than the FEXT crosstalk, because the NEXT does not necessarily propagate through the line length and thus does not experience a propagation loss of the signal [9], [13], [17].

### C. Impulse noise signal

In unshielded twisted pairs, various equipment and environmental disturbances such as signaling circuits, transmission and switching gear, electrostatic discharges, lightning surges and so forth can generate an impulse noise. The impulse noise has some reasonably well-defined characteristics. Features of the typical impulse noise can be summarized as follows:

- occurs about 1-5 times per minute (on an average 4 times per minute),
- peak values in the range 2 - 33 mV,
- most of energy concentrated below 40 kHz,
- time duration in the range 30 - 150  $\mu$ s.

Of course, mentioned features don't characterize all possible impulse noise signals. In the simulation model, therefore, characteristics of the impulse noise signal can be randomly varied.

## III. THE ENVIRONMENT OF POWER DISTRIBUTION CABLES

### A. The multipath signal propagation

The PLC transmission channel has a tree-like topology with branches formed by additional wires tapered from the main path and having various lengths and terminated loads with highly frequency-varying impedances in a range from a few ohms to some kilohms. That's why the PLC signal propagation does not only take place along a direct line-of-sight path between a transmitter and a receiver but also additional paths are used for a signal spreading. This multipath scenario must be seriously considered. The simulation model can be simplified if we approximate infinite number of paths by only  $N$  dominant paths and make  $N$  as small as possible. When more transmissions and reflections occur along the path, then the weighting factor will be smaller. When the longer path will be considered, then the signal contribution from this part to the overall signal spreading will be small due to the higher signal attenuation [4], [18].

### B. The signal attenuation

Characteristics of the PLC transmission environment focused on the multipath signal propagation, the signal attenuation, the noise scenario and the electromagnetic compatibility are introduced in [12]. First, we can present basic characteristics of the PLC channel.

A total signal attenuation on the PLC channel consists of two parts: coupling losses (depending on a transmitter design) and line losses (very high and can range from 40 to 100 dB/km). To find a mathematical formulation for the signal attenuation, we have to start with the complex propagation constant

$$\gamma(\omega) = \sqrt{(R + j \cdot \omega L) \cdot (G + j \cdot \omega C)} = \alpha(\omega) + j \cdot \beta(\omega) \quad (7)$$

depending on the primary cable parameters  $R$ ,  $L$ ,  $G$ ,  $C$ . Then, the frequency response of a transmission line  $\mathcal{H}(f)$  (the transfer function) with the length  $l$  can be expressed as follows ( $\mathcal{U}(x)$  is the voltage at the distance  $x$ ):

$$\mathcal{H}(f) = \frac{\mathcal{U}(x=l)}{\mathcal{U}(x=0)} = e^{-\gamma(f) \cdot l} = e^{-\alpha(f) \cdot l} e^{-j \cdot \beta(f) \cdot l} \quad (8)$$

Considering frequencies in the megahertz range, the resistance  $R$  per length unit is dominated by the skin effect and thus is proportional to  $\sqrt{f}$ . The conductance  $G$  per length unit is mainly influenced by a dissipation factor of the dielectric material (usually PVC) and therefore proportional to  $f$ . With typical geometry and material properties, we can suppose  $G \ll \omega C$  and  $R \ll \omega L$  in the frequency range of interest. Then, cables can be regarded as low loss ones with real valued characteristic impedances and a simplified expression for the complex propagation constant  $\gamma$  can be introduced

$$\gamma(f) = k_1 \cdot \sqrt{f} + k_2 \cdot f + j \cdot k_3 \cdot f = \alpha(f) + j \cdot \beta(f) \quad (9)$$

where constants  $k_1$ ,  $k_2$  and  $k_3$  are parameters summarizing material and geometry properties. Based on these derivations and an extensive investigation of measured frequency responses, an approximating formula for the attenuation factor  $\alpha(f)$  is found in a form

$$\alpha(f) = a_0 + a_1 \cdot f^k \quad (10)$$

that is able to characterize the attenuation of typical power distribution lines with only three parameters, being easily derived from the measured transfer function [18]. Now the propagation loss  $L_{dB}$  is given at the length  $l$  and the frequency  $f$  as

$$L_{dB}(l, f) = -20 \cdot \log_{10} |\mathcal{H}(l, f)| = \frac{20}{\ln 10} \cdot l \cdot \alpha(f) = \frac{20}{\ln 10} \cdot l \cdot (a_0 + a_1 \cdot f^k) \quad [Np]$$

$$\approx 8,686 \cdot l \cdot (a_0 + a_1 \cdot f^k) \quad [dB] \quad (11)$$

We can see a linear dependence of the propagation loss  $L_{dB}$  on the line length  $l$ . Parameters  $a_0$ ,  $a_1$  and  $k$  are characterized by measurements of the transfer function  $\mathcal{H}(f)$  that is much easier than the measurement of primary line parameters  $R$ ,  $L$ ,  $C$ ,  $G$ . If we now merge a signal spreading on all paths together (we can use a superposition), we can receive an expression for the frequency response  $\mathcal{H}(f)$  in a form

$$\mathcal{H}(f) = \sum_{i=1}^N g_i \cdot a(l_i, f) \cdot e^{-j \cdot 2 \cdot \pi \cdot f \cdot \tau_i} \quad (12)$$

where  $a(l_i, f)$  is the signal attenuation proportioned with the length and the frequency and  $N$  is the number of paths in the transmission channel. The delay  $\tau_i$  of the transmission line can be calculated from the dielectric constant  $\epsilon_r$  of insulating materials, the light speed  $c$  and the line length  $l_i$  as follows

$$\tau_i = \frac{l_i \cdot \sqrt{\epsilon_r}}{c} \quad (13)$$

### C. The noise scenario

Unfortunately, in a case of the PLC environment, we can't stay only with the additive white Gaussian noise. The noise scenario is much more complicated, since five general classes of noise can be distinguished in power distribution line channels. These five classes are:

1. *Colored background noise* – caused by a summation of numerous noise sources with low powers. Its PSD varies with the frequency in a range up to 30 MHz (significantly increases toward to lower frequencies) and also with the time in terms of minutes or even hours.
2. *Narrowband noise* – caused by ingress of broadcasting stations. It is generally varying with daytimes and consists mostly of sinusoidal signals with modulated amplitudes.
3. *Periodic impulsive noise asynchronous with the main frequency* – caused by rectifiers within DC power supplies. Its spectrum is a discrete line spectrum with a repetition rate in a range between 50 and 200 kHz.
4. *Periodic impulsive noise synchronous with the main frequency* – caused by power supplies operating synchronously with the main cycle. Its PSD is decreasing with the frequency and a repetition rate is 50 Hz or 100 Hz.
5. *Asynchronous impulsive noise* – caused by impulses generated by the switching transients' events in the network. It is considered as the worst noise in the PLC environment, because of its magnitude that can easily reach several dB over other noise types. Fortunately, the average disturbance ratio is well below 1 percent, meaning that 99 percent of the time is absolutely free of the asynchronous impulsive noise.

The noise types 1, 2 and 3 can be summarized as background noises because they are remaining stationary over periods of seconds and minutes, sometimes even of hours. On the contrary, the noise types 4 and 5 are time-variant in terms of microseconds or milliseconds and their impact on useful signals is much more stronger and may cause single-bit or burst errors in a data transmission [5], [6].

## IV. THE SIMULATION MODEL FOR THE VDSL AND PLC TECHNOLOGIES

For considering of the signal transmission on metallic homogeneous lines by means of the VDSL and PLC technologies, it is necessary comprehensively to know characteristics of negative environmental influences and features of applied modulation techniques. It is difficult to realize of the exact analytical description of complex systems such as the VDSL and PLC systems in the real environment of local access networks. In addition, due to dynamical natures of some processes, it is not suitable. For analyzing of various signal processing techniques used by the VDSL and PLC technologies, a suitable and flexible enough tool are computer simulations and modeling schemes of real environmental conditions at the signal transmission.

For modeling of the VDSL and PLC transmission paths, we used the software program *Matlab* together with additional libraries like *Signal Processing Toolbox* and *Communication Toolbox*. The realized model (Fig. 1) represents the high-speed data signal transmission in both downstream and upstream

directions for the VDSL environment utilizing metallic homogenous lines and for the PLC environment utilizing power distribution lines. This VDSL environment model is the enhanced version of the ADSL environment model introduced [9]. New features of this simulation model are VDSL transmission characteristics and applications of pre-coding techniques and trellis coded modulations. The realized model also represents a high-speed signal transmission in the PLC system utilizing outdoor power distribution lines in downstream and upstream directions [12]. The signal transmission over outdoor power distribution lines represents the transmission between a transmitter in the transformer substation and a receiver in the customer premises.

The requested analysis can be based on computer simulations that cover the most important features and characteristics of the real transmission environment for the VDSL and PLC technologies and result in searching for the optimal combination of advanced modulation, encoding and pre-coding techniques. On Fig. 2, a structure of the VDSL environment block is introduced. Courses of particular PSD noises in this environment realized in the appropriate simulation model are graphically presented on Fig. 3. On Fig. 4, a structure of the PLC environment block is introduced. Courses of particular PSD noises in this environment realized in the appropriate simulation model are graphically presented on Fig. 5.

Basic functional blocks realized in the simulation model are shown on Fig. 1. The VDSL simulation model can be divided into the three main parts:

1. A transmitting part - it is responsible for encoding (using various techniques), for interleaving and for modulating (using various approaches) of signals into a form suitable for the transmission channel.
2. A transmission channel (metallic homogenous lines, power distribution lines) - this part of the model realized appropriate negative influences on the transmitted signal. Above all, it goes about a propagation loss, a signal distortion, crosstalk noises, white and impulse noises, the radio interference for the VDSL environment. For the PLC environment, it goes about the propagation loss, the signal distortion, impulsive, coloured and narrow-band noises. Because these negative influences expressively interfere into the communication path and represent its main limiting factors, they present a critical part of the model and, therefore, it is necessary exactly to recognize and express their characteristics by correct parameters.
3. A receiving part - it is conceptually inverted in a comparison with the transmitter. Its main functions are the signal amplification, demodulation, de-interleaving, removing of the inter-symbol interference and the correction of errored information bits.

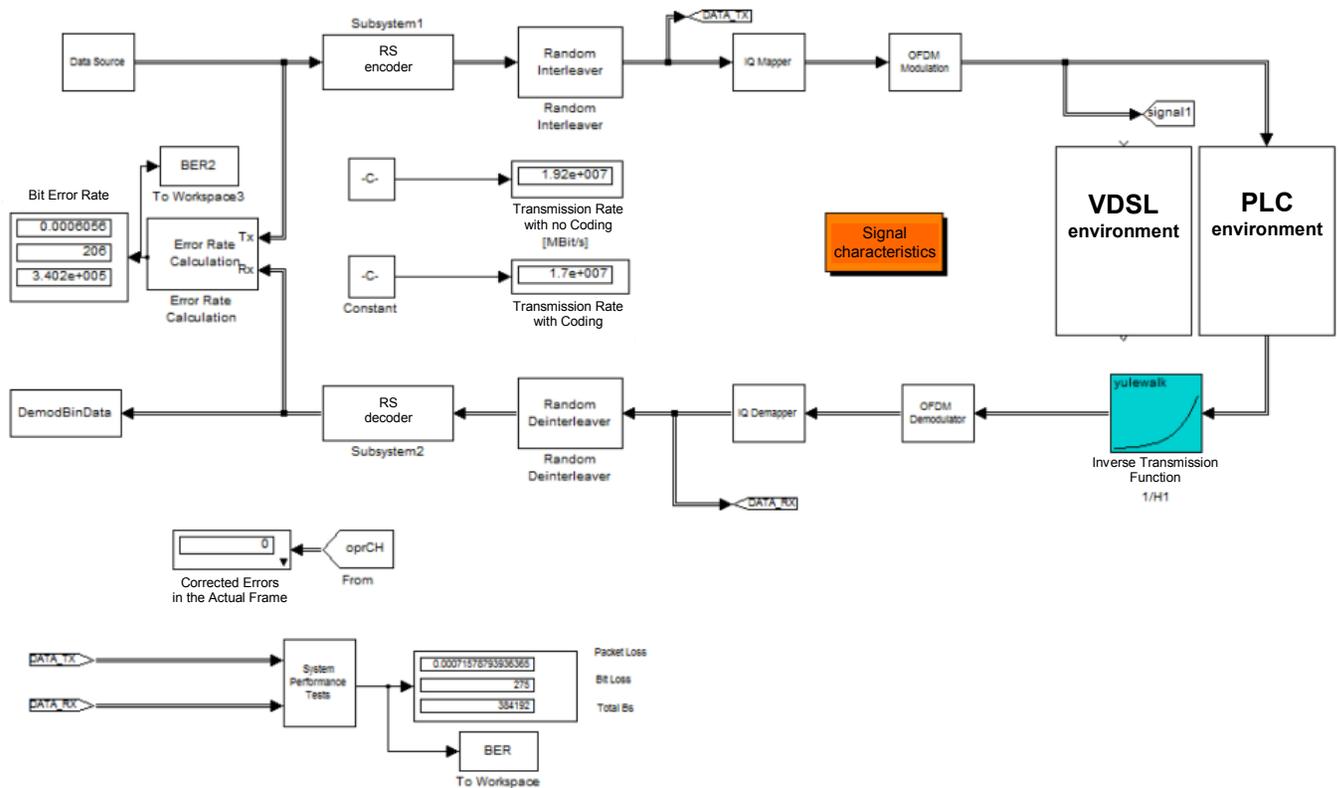


Fig. 1 The Simulink model of VDSL and PLC transmission paths

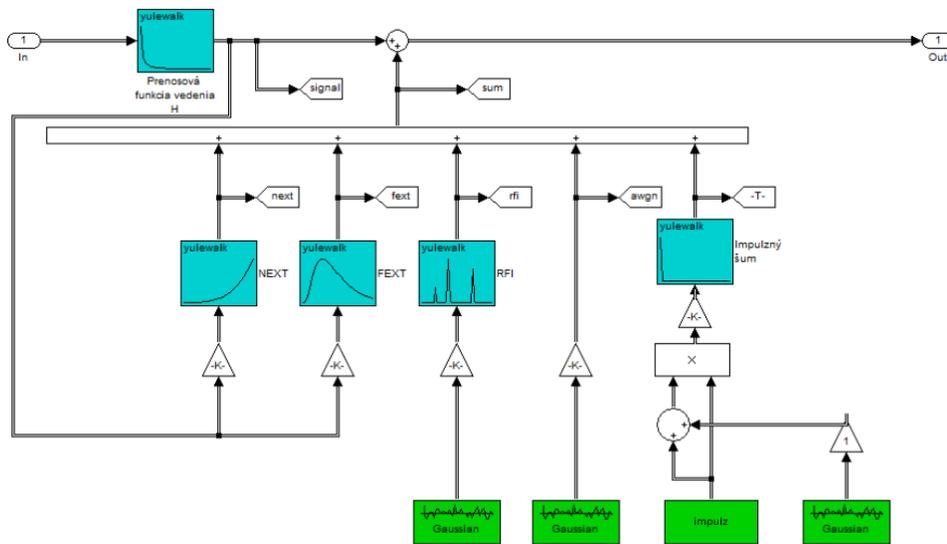


Fig. 2 The VDSL environment block

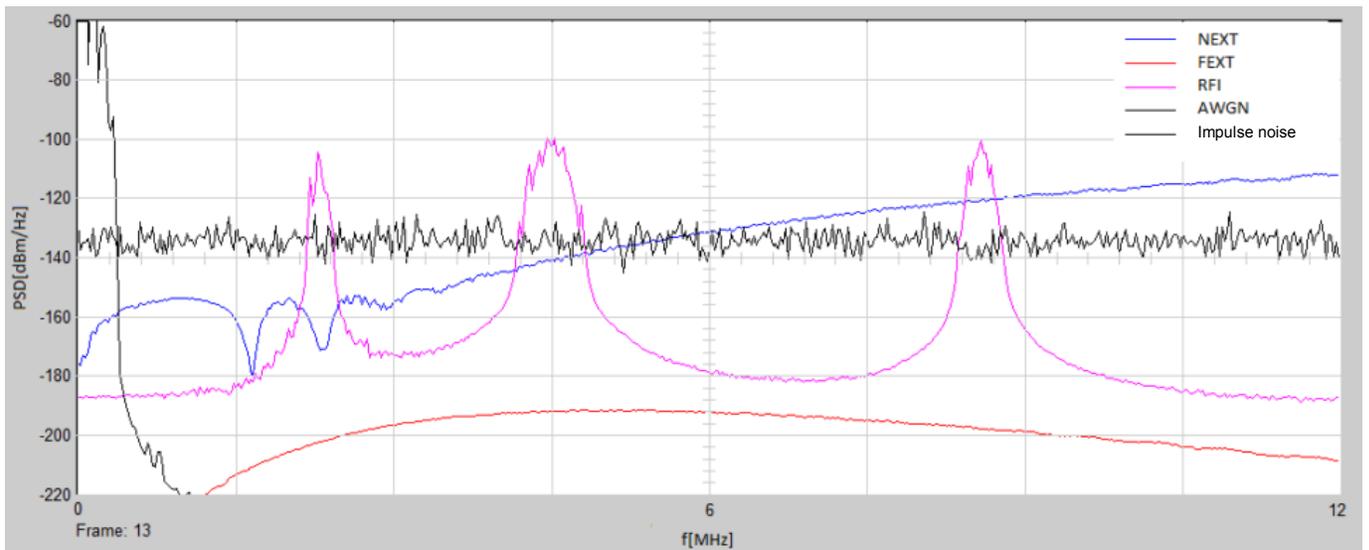


Fig. 3 Particular PSD noises in the VDSL environment

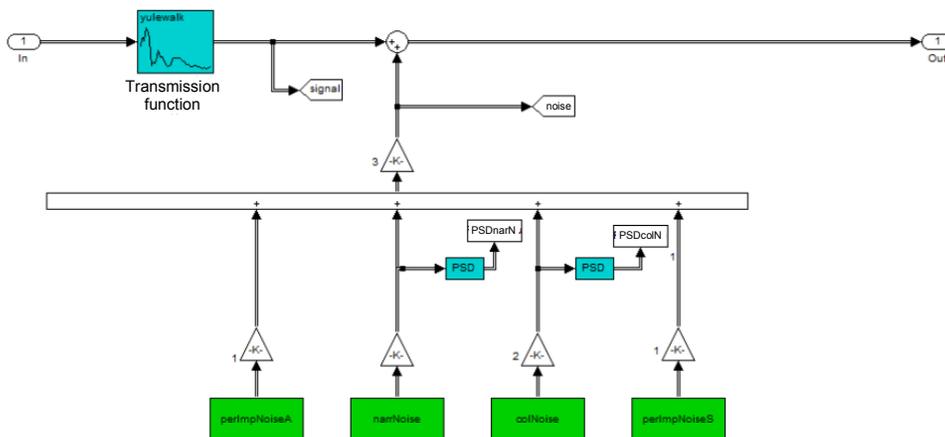


Fig. 4 The PLC environment block

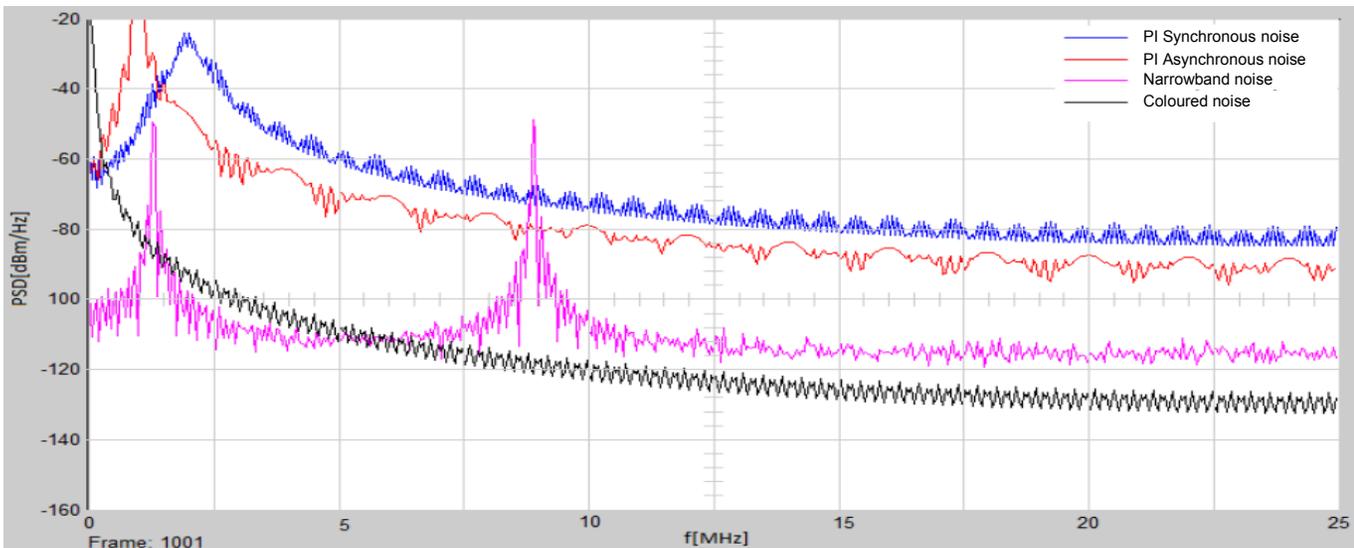


Fig. 5 Particular PSD noises in the PLC environment

## V. THE ENVIRONMENT OF OPTICAL FIBERS

### A. Transmission parameters of the optical fiber

Basic transmission factors of the standard optical single-mode fiber utilized in telecommunications are following:

- linear effects - the attenuation,  
- the dispersion (CD, PMD).
- nonlinear effects - Kerr nonlinearities (FWM, SPM, XPM, XPolM),  
- Scattering nonlinearities (SRS, SBS).

Linear effects represent a majority of losses at the optical signal transmission signal through the optical fiber. These linear effects are mainly caused by the attenuation and the dispersion. The attenuation limits a distance of the optical signal transmission and the dispersion influences transmission rates of optical signals.

Nonlinear effects in the optical fiber may potentially have a significant impact on the performance of WDM optical communication systems. In a WDM system, these effects place constraints on the spacing between adjacent wavelength channels and they limit the maximum power per channel, the maximum bit rate and the system reach [8].

These effects play an important role in a transmission of optical pulses through the optical fiber. We can classify nonlinear effects:

- a) *Kerr nonlinearities*, which is a self-induced effect in that the phase velocity of waves depends on the wave's own intensity. The Kerr effect describes a change in the refractive index of the optical fiber due to an electrical perturbation. Due to the Kerr effect, we are able to describe following effects :
  - Self-Phase Modulation SPM - the effect that changes the refractive index of the transmission media caused by an intensity of the pulse.
  - Four Wave Mixing FWM - the effect, in which mixing of optical waves rise the fourth wave that, can occur in the same wavelength as one of mixed waves.
  - Cross-Phase Modulation XPM - the effect where a wave of the light can be changed by the phase of another wave of the light with different wavelengths. This effect causes a spectral broadening.
- b) *Scattering nonlinearities*, which occur due to an inelastic photon scattering to the lower energy photon. We can say that energy of the light wave is transferred to another wave with different wavelengths. Two effects appear in the optical fiber:
  - Stimulated Brillion and Raman scattering - effects that change a variance of the light wave into different waves when intensity reaches certain threshold.

Knowing which fundamental linear and nonlinear interactions dominate is helpful to conceive techniques that improve a transmission of optical signals, including advanced modulation formats, a digital signal processing and a distributed optical nonlinearity management.

### B. The attenuation

The most important parameter of optical fibers is the attenuation that represents a transmission loss. In practical way, it is a power loss that depends on a length of the transmission path. The attenuation leads to a reduction of the signal power as the signal propagates over some distance. When determining the maximum distance that a signal propagate for a given transmitter power and receiver sensitivity, the attenuation must be considered. The total signal attenuation  $a$  [dB] defined for a particular wavelength can be expressed as

$$a[\text{dB}] = 10 \log_{10} \frac{P_i}{P_o} \quad (14)$$

where  $P_i$  is the input power and  $P_o$  is the output. The attenuation coefficient  $\alpha$  [dB/km] of the optical fiber can be obtained by measuring the input and the output optical power levels. The specific attenuation of the optical fiber along the fiber length  $L$  [km] can be expressed as

$$\alpha[\text{dB/km}] = \frac{10 \log_{10} \frac{P_i}{P_o}}{L} = \frac{a}{L} \quad (15)$$

where  $L$  is the optical fiber's length in [km]. For the link length  $L$ , the  $P(L)$  must be greater than or equal to the receiver sensitivity  $Pr$ .

The attenuation of optical fibers is mainly caused by material absorption losses, radiation scattering and by bending losses. The fiber loss is not only source of the optical signal attenuation along transmission lines. Fiber splices and fiber connectors also cause the signal attenuation. The number of optical splices and connectors depends on the transmission length and must be taken into account unless the total attenuation due to fiber joints is distributed and added to the optical fiber attenuation.

### C. The dispersion

The dispersion is a widening of the pulse duration as it travels through the optical fiber. We distinguished two basic dispersive forms - the intermodal dispersion and the chromatic dispersion. Both cause an optical signal distortion in multimode optical fibers, whereas a chromatic dispersion is the only cause of the optical signal distortion in single-mode fibers.

The chromatic dispersion CD represents a fact that different wavelengths travel at different speeds, even within the same mode. In a dispersive medium, the index of refraction  $n(\lambda)$  is a function of the wavelength. Thus, certain wavelengths of the transmitted signal will propagate faster than other wavelengths. The CD dispersion is the result of material dispersion, waveguide dispersion and profile dispersion.

The chromatic dispersion is caused by different time of the spreading wave through fiber for a different wavelength and it depends on the spectral width of the pulse. As mentioned before, optical fiber represents the transmission system. Then the system has transfer function  $\mathcal{H}_0(\omega)$  given by equation (16). We assume that  $|\mathcal{H}_0(\omega)| = 1$  and we can expand phase into the Taylor series as is given by equation (17). If we consider first two coefficients, then we can write transfer function as given by equation (18).

$$\mathcal{H}_0(\omega) = |\mathcal{H}_0(\omega)| \cdot e^{-j\varphi(\omega)} \quad (16)$$

$$\varphi(\omega) = \left[ \varphi_0 + \frac{d\varphi}{d\omega}(\omega - \omega_0) + \frac{1}{2} \frac{d^2\varphi}{d\omega^2}(\omega - \omega_0)^2 + \frac{1}{6} \frac{d^3\varphi}{d\omega^3}(\omega - \omega_0)^3 \dots \right] \quad (17)$$

$$\mathcal{H}_0(\omega) = e^{-j\varphi_0} \cdot e^{-j\frac{d\varphi}{d\omega}(\omega - \omega_0)} \cdot e^{-j\frac{d^2\varphi}{d\omega^2}(\omega - \omega_0)^2} \quad (18)$$

where  $\mathcal{H}_0(\omega)$  is a transfer function,  $\varphi_0$  is an initial phase of the system and  $\omega_0$  is an initial angular frequency

After few operations, we can obtain time  $t$  from the transfer function, which represents the travel time of the pulse through the fiber, the signal phase shift  $\Delta\varphi$  and the Group Velocity Dispersion GVD coefficient. These parameters are described by two equations:

$$t = \frac{1}{2\pi} \frac{d\varphi}{df_m} \quad (19)$$

$$GVD = \frac{1}{2\pi} \frac{d^2\varphi}{df_m^2} \quad (20)$$

The chromatic dispersion causes broadening and phase changing of the signal. Then pulses at the end of optical fibers may start to overlap and this effect is called as the Inter Symbol Interference ISI.

The polarization mode dispersion PMD is another complex optical effect that can occur in optical single-mode fibers. The SMF support two perpendicular polarizations of the original transmitted signal. If a fiber is not perfect, these polarization modes may travel at different speeds and, consequently, arrive at the end of the fiber at different times. The difference in arrival times between the fast and slow mode axes is the PMD. Like the CD, the PMD causes digitally-transmitted pulses to spread out as the polarization modes arrive at their destination at different times.

$$\Delta\tau = D_{PMD} \cdot \sqrt{L} \quad (21)$$

The main problem with the PMD in optical fiber systems is its stochastic nature, letting the principal state of polarization PSP and the differential group delay DGD vary on timescales between milliseconds and months.

The resulting overall dispersion is composed of chromatic dispersion and polarization mode dispersion path and is given by the resulting relation [10], [11]

$$D = \sqrt{D_{CD}^2 + D_{PMD}^2} \quad (22)$$

#### D. The Four Wave Mixing effect

The four wave mixing FWM is a parametric interaction among waves satisfying a particular phase relationship called the phase matching. This nonlinear effect occurs only in systems that carry more wavelengths through the optical fiber and it is classified as a third-order distortion phenomenon. In this case, we are assuming that three linearly polarized monochromatic waves with angular frequencies  $\omega_j$  ( $j = 1, 2, 3$ ) are propagating. If we consider third-order polarization vector  $\mathbf{P}$  given by equation (23) that characterizes the medium and it is a function of the electrical field, and simplified it, we obtain his components: three components have the frequencies of the input field, the others have frequencies  $\omega_k$  given by equation (24)

$$\bar{P} \approx \varepsilon_0 \left\{ \chi^{(1)} \bar{E} + \chi^{(2)} : \bar{E}\bar{E} + \chi^{(3)} : \bar{E}\bar{E}\bar{E} \right\} \quad (23)$$

where  $\chi^{(1)}$  is the linear susceptibility,  $\chi^{(2)}$ ,  $\chi^{(3)}$  is the second- and the third-order susceptibility and  $\mathbf{E}$  represent vector of electrical field of mode.

$$\omega_k = \omega_1 \pm \omega_2 \pm \omega_3 \quad (24)$$

As we can see from the equation (9), nonlinear interaction generates new frequency components of the material polarization vector, which can interfere with input fields if a phase matching condition is obtain. The most frequency components fall away from our original bandwidth or near it. Frequency components that directly overlap with bandwidth will cause an interference with original waves.

The power of new generated waves can be obtain by solving coupled propagation equations of four interacting waves. We assume that the new generated FWM wave is mainly depended on three nearest waves of the light, so the power  $A_k^2$  at the frequency  $\omega_k = \omega_1 + \omega_2 - \omega_3$  is given by

$$A_k^2 = 4\eta\gamma^2 d_e^2 L_e^2 A_1^2 A_2^2 A_3^2 e^{-\alpha l} \quad (25)$$

where factor  $\eta$  is the FWM efficiency,  $\gamma$  is the nonlinear coefficient,  $L_e$  is the effective length,  $A_1^2(z)$ ,  $A_2^2(z)$ ,  $A_3^2(z)$  are powers of input waves,  $l$  is the fiber length,  $\alpha$  is the attenuation and  $d_e$  the so-called degeneracy factor (equal to 3 if the degenerative FWM is considered, 6 otherwise).

The power of the FWM represents sum of the partial power from interacting waves, which degenerate the signal. This power of the FWM is different for each channel and change with the parameter of interacting signals.

As we can see from the equation (25), the nonlinear effect FWM is mainly rising with increasing powers of interacting signals and the shape of the FWM effect depends on the modulation and the bit rate of these signals. If input powers of signals are too high, the scattering nonlinearities occur and transmission would not be possible. However, the scattering nonlinearities are not presented. The power also depends on the channel spacing and on the dispersion. If we use negative dispersion fibers, the FWM effect will be more intensive and the SNR will fall to values unsuitable to transmit. If we used standard fibers, we can decrease the FWM effect, but we cannot use high bit rates due the dispersion [2], [10], [11].

#### E. The Self-Phase Modulation effect

The self-phase modulation SPM has an important impact on high data speed communication systems that use the dense wavelength division multiplexing. The SPM effect occurs due to the Kerr effect in which the refractive index of optical fiber increases with the optical intensity decreasing the propagation speed and thus inducts the nonlinear phase shift. The relation between intensity and refractive index can be described

$$n_r = n_{r0} + \bar{n}_2 I(t) \quad (26)$$

where  $n_r$  is the refractive index dependent on intensity,  $n_{r0}$  is the linear refractive index,  $\bar{n}_2$  is the nonlinear refractive index and  $I(t)$  is the intensity.

This varying parameter  $n_r$  causes the SPM effect in which the signal phase propagating through the optical fiber changes with the distance and can be described by

$$\phi = \left( n_0 z + \phi_0 \right) + \frac{2\pi}{\lambda} \bar{n}_2 I(t) z \quad (27)$$

where  $\phi_0$  represents the initial phase.

The equation (27) shows that the different phase shift occurs during the pulse propagation caused by the intensity dependence of phase fluctuations. This variation in phase with time is responsible for changes in frequency spectrum by following equation

$$\omega = \frac{d\phi}{dt} \quad (28)$$

If we assume the variation of phase with intensity pulse then we can write following equation

$$\omega' = \omega_0 + \frac{d\phi}{dt} \quad (29)$$

where  $\omega'$  is the signal frequency affected with the SPM effect,  $\omega_0$  is the initial signal frequency. This variation in signal frequency is called the frequency chirp.

The SPM effect impact becomes more significant with increasing an optical fiber length, especially when optical amplifiers such as EDFA and RAMAN are used [3], [10], [11].

*F. The Cross-Phase Modulation Effect XPM*

The Cross-phase modulation (XPM) is very similar to the SPM in which the intensity from different wavelength channels changes the signal phase and thus the XPM occurs only in WDM systems. In fact, the XPM converts power fluctuations in a particular wavelength channel to phase fluctuations in other co-propagating channels. The XPM effect results to spectral broadening and distortion of the pulse shape.

If we assume  $N$  signal having different carrier frequencies propagating in an optical fiber, the nonlinear signal phase depends on signal intensities at different frequencies. This phase shift can be described by expression

(30)

$$\Delta\phi_i = \frac{2\pi n_2 z}{\lambda} \left[ I_i(t) + 2 \sum_{i \neq j} I_j(t) \right]$$

where the first term in bracket represents the SPM effect and the second term represents XPM effect.

In the equation (30), the factor 2 has its origin in a form of the nonlinear susceptibility and represents the XPM twice as effective as the SPM for the same power amount. The XPM effect affects the signal only the interacting signals superimpose in time. The XPM effect can decrease a system performance even greater than the SPM effect, especially in case of 100 channels systems [3], [10], [11].

VI. THE SIMULATION MODEL FOR THE OPTICAL COMMUNICATIONS

For modeling of the optical transmission path, we used the software program *Matlab 2010 Simulink* together with additional libraries like *Communication Blockset* and *Communication Toolbox*. The realized model (Fig. 6) represents the signal transmission in the environment utilizing optical fibers for very high-speed data signals in both directions. Optical communication technologies will always be facing the limits of high-speed signal processing and modulation, which is an important factor to take into account when discussing advanced optical modulation formats. The main task of the simulation model is an analysis of various modulation and encoding techniques.

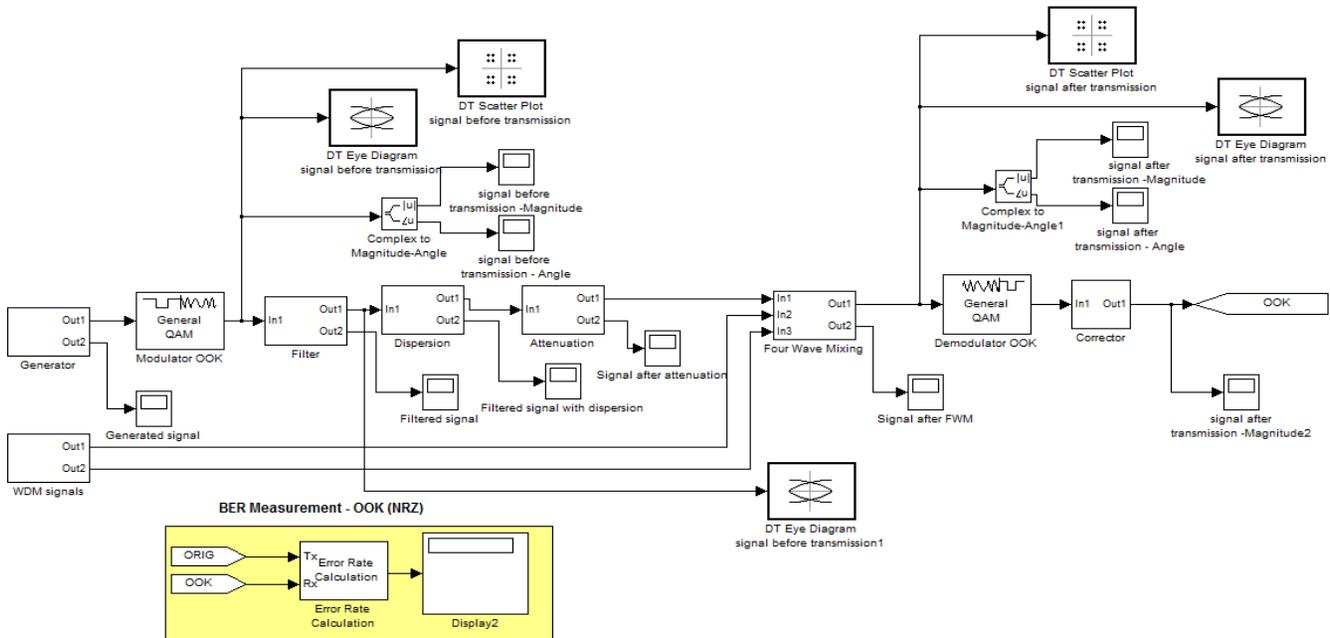


Fig. 6 The Simulink model of the optical transmission path

Basic functional blocks realized in the optocommunication simulation model can be divided into the three main parts:

1. A transmitting part - it is responsible for the generating and for modulating of generated signals according to required information inputs into a form suitable for the transmission channel. The modulation block contains a basic set OOK modulation and other variations of modulation techniques.
2. A transmission channel (the optical fiber) - this part of the model realized negative influences on the transmitted signal. Because these negative influences expressively interfere into the communication and represent its main limiting factors, they present a critical part of the model and, therefore, it is necessary exactly to recognize and express their characteristics by correct parameters.
3. A receiving part - it is conceptually inverted in a comparison with the transmitter. At the receiver side, a signal is demodulated by appropriate demodulator and the BER ratio is calculated. Also, blocks for graphical presenting of transmitted optical signals can be utilized.

## VII. CONCLUSION

The first part of the lecture analyzes basic features of the real transmission environment of metallic homogeneous lines and presents possibilities for modeling and simulating of the information signal transport in this environment by means of the VDSL technology. We focused on the determination and the analysis of concrete characteristic features for substantial negative influences of internal and external environments and on the representation of frequency dependencies of transmitted VDSL signals.

The second part of the lecture analyzes basic features of the real transmission environment of the outdoor power distribution lines and presents possibilities for modeling and simulation of the information signal transmission in this environment by means of the PLC technology. We focused on transmission characteristics of the PLC channel, namely the multipath signal propagation, the signal attenuation and the interference scenario revealing different classes of the impulsive noise. We created a model of the complex frequency response in a range from 500 kHz up to 30 MHz. Moreover, we realized experimental measurements for verification of the parametric model for reference channels.

Basic features and characteristics of negative environmental influences at the signal transmission in VDSL and PLC environments can be used for modeling spectral characteristics of signals on the transmission path. The VDSL and PLC simulation models allow determining main problems that can arise at the VDSL or PLC signal transmission. For realizing of individual model blocks, we concentrated on the choice of appropriate parameters so that these blocks could be adjusted and modified for future demands. The PLC simulation model is verified by measurements in the real PLC transmission environment that confirmed its satisfactory conformity with real transmission conditions.

The fourth part of the lecturer analyzes transmission parameters for the optical transmission medium and presents possibilities for modeling and simulation of the information signal transmission in the environment of optical single-mode fibers. We focused on linear transmission factors – the attenuation and the dispersion - and on nonlinear effects – the four wave mixing, the self-phase modulation and the cross-phase modulation. Nonlinear effects in the optical fiber may potentially have a significant impact on the performance of WDM optical communication systems.

The simulation model for the optical communication path represents the signal transmission in the optical environment for very high-speed data signals in both directions. Knowing which fundamental linear and nonlinear interactions dominate in the optical transmission medium is helpful to conceive techniques that improve a transmission of optical signals, including advanced modulation formats, encoding techniques, digital signal processing and a distributed optical nonlinearity management.

## REFERENCES

- [1] Binh, L.N., *Optical Fiber Communications Systems*, CRC Press, ISBN 978-1-4398-0620-3, Boca Raton, United States of America, 2010.
- [2] F. Čertík, R. Róka, "The Nonlinear FWM Effect and its Influence on Optical Signals Utilized Different Modulation Techniques in the WDM Transmission Systems", In: OK 2012 - 24th Conference, Praha (Czech Republic), 2012, ISBN 978-80-86742-36-6.
- [3] F. Čertík, R. Róka, "Nonlinear SPM and XPM Effects and their Influence on Optical Signals Utilized Different Modulation Techniques in WDM Transmission Systems", OK 2014 - 26th Conference, Praha (Czech), 2014, ISBN 978-80-86742-39-7.
- [4] Ferreira, H.C.; Lampe, L.; Newbury, J.; Swart, T.G., *Power Line Communications*, John Wiley & Sons, ISBN 978-0-470-74030-9, Chichester, United Kingdom, 2010.
- [5] Götz, M.; Rapp, M.; Dostert, K., "Power Line Channel Characteristics and Their Effect on Communication System Design", *IEEE Communications Magazine*, Vol.42, No.4, April 2004, pp. 78-86, ISSN 0163-6804.
- [6] Held, G., *Understanding Broadband over Power Line*, Auerbach Publications, ISBN 0-8493-9846-0, Boca Raton, United States of America, 2006.
- [7] Kaminow, I.P.; Li, T.; Willner, A.E., *Optical Fiber Telecommunications V B: Systems and Networks*, Elsevier Inc., ISBN 978-0-12-374172-1, San Diego, United States of America, 2008.
- [8] Mukherjee, B., *Optical WDM Networks*, Springer Science+Business Media Inc., ISBN 978-0387-29055-3, New York, United States of America, 2006.
- [9] Róka, R.; Cisár, R., "The Analysis of Negative Influences in the Environment of Homogeneous Symmetric Lines at the Signal Transmission by Means of the ADSL Technology", *Journal of Electrical Engineering - EČ*, Vol.53, No.9-10, September 2002, pp. 241-249, ISSN 1335-3632.

- [10] R. Róka, F. Čertík, "Modeling of Environmental Influences at the Signal Transmission in the Optical Transmission Medium", *International Journal of Communication Networks and Information Security*, 2012, Vol. 4, No. 3, pp. 146-162, ISSN 2073-607X.
- [11] Róka, R., Čertík, F., *Simulation Tools For Broadband Passive Optical Networks*, In: Simulation Technologies in Networking and Communications: Selecting the Best Tool for the Test, CRC Press, ISBN 978-1482225495, Boca Raton, USA, November 2014.
- [12] Róka, R., Dlháň, S., "Modeling of transmission channels over the low-voltage power distribution network", *Journal of Electrical Engineering - EČ*, Vol.56, No.9-10, September 2005, pp. 237-245, ISSN 1335-3632.
- [13] Róka, R., "Environmental Influences on the Power Spectral Densities of VDSL Signals", *Journal of Electrical Engineering - EČ*, Vol.55, No.1-2, January 2004, pp. 18-24, ISSN 1335-3632.
- [14] Róka, R., "Modeling of Environmental Influences at the Signal Transmission by means of the VDSL and PLC Technologies", *International Journal of Electrical Communication Networks and Information Security - IJCNIS*, Vol. 1, No. 1, April 2009, pp. 6-13, ISSN 2073-607X.
- [15] Róka, R., *Fixed Transmission Media*. In: Technology and Engineering Applications of Simulink, InTech, ISBN 978-953-51-0635-7, Rijeka, Croatia, May 2012.
- [16] Róka, R., "Analysis of Advanced Modulation Techniques in the Environment of Metallic Transmission Media", In: TSP 2014 - 37th International Conference on Telecommunications and Signal Processing, Berlin (Germany), 1. - 3. 7. 2014, pp. 78-84, ISBN 978-80-214-4983-1, ISSN 1805-5435
- [17] Werner, J. J., "The HDSL Environment", *IEEE Journal on Selected Areas in Communications*, Vol.SAC-9, No.6, August 1991, pp. 785-800, ISSN 0733-8716.
- [18] Zimmermann, M.; Dostert, K., "Multipath Model for the Powerline Channel", *IEEE Transactions on Communications*, Vol.50, No.4, April 2002, pp. 553-559, ISSN 0090-6778.

## ABBREVIATIONS

|       |                                  |
|-------|----------------------------------|
| ADSL  | Asymmetric DSL                   |
| CD    | Chromatic Dispersion             |
| DGD   | Differential Group Delay         |
| FEXT  | Far-End Crosstalk                |
| FWM   | Four Wave Mixing                 |
| GVD   | Group Velocity Dispersion        |
| ISI   | Inter-Symbol Interference        |
| NEXT  | Near-End Crosstalk               |
| PLC   | Power Line Communication         |
| PMD   | Polarization Mode Dispersion     |
| PON   | Passive Optical Network          |
| PSD   | Power Spectral Density           |
| SBS   | Stimulated Brillouin Scattering  |
| SPM   | Self-Phase Modulation            |
| SRS   | Stimulated Raman Scattering      |
| VDSL  | Very high bit rate DSL           |
| WDM   | Wavelength Division Multiplexing |
| XPM   | Cross-Phase Modulation           |
| XPolM | Cross-Polarization Modulation    |

## AUTHOR BIOGRAPHY

Rastislav Róka (Assoc. Prof.) was born in Šaľa, Slovakia on January 27, 1972. He received his MSc. and PhD. degrees in Telecommunications from the Slovak University of Technology, Bratislava, in 1995 and 2002. Since 1997, he has been working as a senior lecturer at the Institute of Telecommunications, FEI STU, Bratislava. Since 2009, he is working as an associated professor at this institute. His teaching and educational activities are realized in areas of fixed transmission media, digital and optocommunication transmission systems and network. At present, his research activity is focused on the signal transmission through optical transport, metropolitan and access networks by means of new WDM and TDM technologies using advanced optical signal processing included various modulation and coding techniques and through metallic access networks by means of xDSL, HFC and PLC technologies. His main effort is dedicated to effective utilization of the optical fiber's transmission capacity of the broadband passive optical networks by means of DBA and DWA algorithms applied in various advanced hybrid optical network infrastructures.

# Some recent advances of ultrasonic diagnostic methods applied to materials and structures (including biological ones)

L. Nobile, S. Nobile

**Abstract**— This paper gives an overview of some recent advances of ultrasonic methods applied to materials and structures (including biological ones), exploring the broad applications of these emerging inspection technologies to civil engineering and medicine. . In confirmation of this trend, some results of an experimental research carried out involving both destructive and non-destructive testing methods for the evaluation of structural performance of existing reinforced concrete (RC) structures are discussed in terms of reliability. As a result, Ultrasonic testing can usefully supplement coring thus permitting less expensive and more representative evaluation of the concrete strength throughout the whole structure under examination.

**Keywords**—diagnostics, nondestructive test, structural performance, ultrasound, medicine, ultrasonic method.

## I. INTRODUCTION

ULTRASONIC method is a form of Non-Destructive Testing performed in Engineering for the inspection without damaging the parts or components and for the characterization of materials. The advantages of this method include flexibility, low cost, in-line operation, and providing data in both signal and image formats for further analysis. In Engineering, ultrasonic testing is often performed on steel and other metals and alloys, on concrete, wood, plastics, ceramics and composites. The main applications in Medicine are diagnostics and therapy of conditions/diseases involving most organs of the body

Since the 1940s ultrasonic test instruments have been employed to detect hidden cracks, voids, porosity, and other internal discontinuities as well as to measure thickness and analyze material properties. The main advantages are: high sensitivity in detecting small flaws; ease of performance; greater accuracy than other nondestructive methods in determining the depth of internal flaws. The main disadvantages are: variability due to operator-dependency; technical limitations regarding rough, irregular, very small, thin or not homogeneous parts; need for surface preparation

by cleaning and removing loose scale, paint, etc. (although paint that is properly bonded to a surface does not need to be removed).

Many international standards covering ultrasonic testing methods in Engineering are published by ISO (International Organization for Standardization) and by CEN (European Committee for Standardization). A standard is a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose

Some of the latest standards are:

- ISO 17405:2014, Non-destructive testing -- Ultrasonic testing -- Technique of testing claddings produced by welding, rolling and explosion
- ISO 12715:2014, Non-destructive testing -- Ultrasonic testing -- Reference blocks and test procedures for the characterization of contact probe sound beams
- ISO 16809:2012, Non-destructive testing -- Ultrasonic thickness measurement
- ISO 16810:2012, Non-destructive testing -- Ultrasonic testing -- General principles
- ISO 16811:2012, Non-destructive testing -- Ultrasonic testing -- Sensitivity and range setting
- ISO 16826:2012, Non-destructive testing -- Ultrasonic testing -- Examination for discontinuities perpendicular to the surface
- ISO 16827:2012, Non-destructive testing -- Ultrasonic testing -- Characterization and sizing of discontinuities.
- ISO 16831:2012 , Non-destructive testing -- Ultrasonic testing -Characterization and verification of ultrasonic thickness measuring equipment.

In 1942, Floyd Firestone [1] patented an instrument he called the Supersonic Reflectoscope, which is generally regarded as the first practical commercial ultrasonic flaw detector that uses the pulse/echo technique commonly employed today. It led to the development of many commercial instruments that were introduced in the following years.

*“My invention pertains to a device for detecting the*

Prof. L. Nobile is with the Dept. of Civil, Chemical, Environmental, and Materials Engineering (DICAM) of the University of Bologna-Campus of Cesena, via Cavalcavia 61, 47521 Cesena, ITALY (phone: +390547338311; fax: +390547338307; e-mail: lucio.nobile@unibo.it).

Dr. S. Nobile, PhD, is with the Maternal and Child Department, Ospedali Riuniti di Ancona, Italy , via F. Corridoni 11,60123 Ancona, ITALY (E-mail: stefano.nobile@ospedaliriuniti.marche.it).

*presence of inhomogeneities of density or elasticity in materials. ....My device may also be used for the measurement of the dimensions of objects, and is particularly useful in those cases where one of the faces to which the measurement extends is inaccessible. ....The general principle of my device consists in the sending of high frequency vibrations into the part to be inspected, and the determination of the time intervals of arrival of the direct and reflected vibrations at one or more stations on the surface of the part....”.*

This pioneer patent has been cited in about 111 other patents.

In the late 1940s, researchers in Japan pioneered the use of ultrasonic testing in medical diagnostics using early B-scan equipment that provided a two-dimensional profile image of tissue layers. By the 1960s, early versions of medical scanners were introduced to diagnose tumors, gallstones, and similar conditions.

The latest advances in ultrasonic instruments have been based on the digital signal processing techniques and the inexpensive microprocessors that became available from the 1980 onwards. This has led to the latest generation of miniaturized, highly reliable portable instruments and on-line inspection systems for flaw detection, thickness gaging, and acoustic imaging.

The aim of this paper is to give an overview of some advanced ultrasonic diagnostic methods applied to materials and structures (including biological ones), exploring the broad applications of these emerging inspection technologies to civil engineering and medicine. In confirmation of this trend, some results of an experimental research carried out involving both destructive and non-destructive testing methods for the evaluation of structural performance of existing reinforced concrete (RC) structures are discussed in terms of reliability.

## II. ADVANCES OF ULTRASOUNDS IN MEDICINE

The introduction of ultrasonic methods in Medicine in the late '60s is one of the most important innovations of the last decades.

Most UltraSound (US) machines include a transducer array, a beam-former, a processor, and a display, and use electroacoustic transducers, which convert electrical energy into mechanical energy and vice versa. The beam-former sets the phase delay and amplitude of each transducer element to enable dynamic focusing and beam steering. Where appropriate, a lens is mounted on the transducer array to focus the transmitted pulses and received echoes. In operation, the transducer array directs a number of pulses towards the anatomical area of a patient to be imaged, and after a variable propagation delay receives echoes that are reflected back by the patient's anatomical structures.

The ultrasound beam is attenuated by the organs and tissues (absorption), and this phenomenon increases with the viscosity and density of the biological structures as well as with the frequency of the ultrasound. Thus, the higher the

frequency of ultrasound, the better is the resolution attained; however, the penetration of the ultrasound into the body will be less, and deep structures will be poorly investigated. In practice, different ranges of frequency are used for examination of different parts of the body: 3–5 MHz for abdominal areas, 5–10 MHz for small and superficial parts and 10–30 MHz for the skin or the eyes. The received signal can then be presented on a display for immediate examination or recorded for a later review. Moreover, computerized image processing may improve image quality.

In medical practice, the most used modalities of signal display are three: M-mode, which shows the ranges of targets along one scan line versus time; B-mode, which provides a cross-sectional image of the body, built up by sweeping a beam sideways through a chosen scan plane; and Doppler ultrasound, which is used to study blood flow as scattering blood cells move towards or away from the probe producing the Doppler effect.[2-4]

The Doppler effect can be employed to study the movement of blood, and consequently to perform a detailed functional assessment of the cardiovascular system as well as evaluation of inflammatory disorders (i.e. intestinal US for inflammatory bowel disease). The injection of intravenous contrast agents, microbubbles, improves the visibility of small vessels with color Doppler; therefore, contrast agents allow a more detailed image of the vascularity of organs (which in some cases is an expression of inflammation) or tumours.

A novel processing modality, three-dimensional imaging, has been recently patented by Angelsen and Johansen [5] in order to allow a better representation of anatomic structures during placement of devices in the heart, guidance of electrophysiology ablation, or guidance in minimal invasive surgery.

Biologic advantages of ultrasounds are ease of use and lack of radiation exposure and carcinogenic properties; disadvantages are the operator-dependency effectiveness and - even if rare- the risk of heating, cavitation and direct damage of cells and organs.

Ultrasounds can be used either for diagnostic and interventional/therapeutic purposes, with appropriate devices and in support of invasive and surgical procedures to increase efficacy and reduce complications (i.e. placement of intravascular catheters, performance of biopsy).

Regarding diagnostics, US devices are increasingly used as a non-invasive imaging tool to evaluate anatomy and detect a wide range of diseases and conditions in all age groups: every anatomical system and apparatus can be studied, and US-based functional studies have greatly reduced the use of radiations and invasive procedures. The size, shape, echo pattern, vascularity and position of organs and other structures can be demonstrated.

Thanks to their biological safety, low cost, lack of radiation exposure and carcinogen properties, US are currently adopted in prenatal medicine, childhood and adulthood. Ultrasound probes can be applied over the skin (transcutaneously) or

internally (i.e. transesophageal, rectal, vaginal US) for a better representation of internal organs.

The main structures which can be evaluated are listed by site in Table 1. A typical ultrasound image of the heart, often referred to simply as echocardiogram, is reported in Fig.1.

Table 1: Main sites and organs which can be evaluated by ultrasounds.

|                    |   |
|--------------------|---|
| Head               | Brain hemorrhage, infarctions, edema, congenital abnormalities (in small infants through bone openings: fontanelles); eye; salivary glands  |
| Neck               | Thyroid and parathyroid glands, lymph nodes, abscesses, vessels   |
| Chest              | Chest wall, pleura, lung (peripheral areas), mediastinum, heart and great vessels   |
| Abdomen and pelvis | Gastrointestinal system (intestine, liver, gallbladder, pancreas), genitourinary apparatus (kidneys, ureters, bladder, uterus, salpinges, ovaries, prostate), spleen, adrenal glands, fluidcontaining structures (cysts, cancer), great vessels and lymph nodes |
| Scrotum            | Testicles, tumors, hernias  |
| Extremities        | Joints, muscles and connective tissue, vessels  |

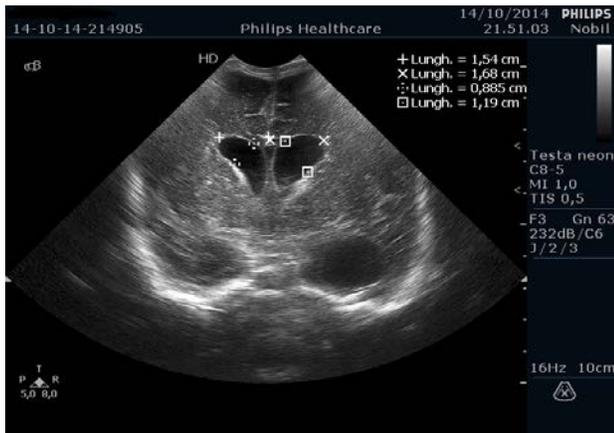


Fig.1 A Typical echocardiogram

Regarding procedures and therapeutic use, US can be employed directly (i.e. breaking of stones) or indirectly (i.e. US-guided procedures, release of drugs)[6]. A list of diagnostic and therapeutic procedures is shown in Table 2.

The evaluation and management of normal and problem pregnancy is one of the most important advances achieved by ultrasounds in the last decades. US enables the detection of the number of the fetuses, position, organ development and

congenital abnormalities, placenta, umbilical cord blood flow. In recent years interventional US is being increasingly used for diagnostic and therapeutic procedures, including fetal noninvasive surgery. In particular, amniocentesis and chorionic villus sampling allow the detection of genetic abnormalities; amnioreduction and amnioinfusion are used to increase or decrease the volume of the amniotic fluid when alterations in its production occur; fetal blood sampling and intrauterine fetal transfusion allow the detection and treatment of blood disorders.

Regarding other recent applications, a patent by Palmeri and al. [7] introduced a method to evaluate liver stiffness (expression of fibrosis and cirrhosis) by means of shear waves which are emitted and detected by an ultrasound transducer, thus limiting the need for liver biopsy - a procedure which could be associated with significant risk of bleeding, particularly among patients at high risk (i.e. liver cirrhosis).

Table 2: diagnostic and therapeutic procedures performed with ultrasounds

|                        |  |
|------------------------|--|
| Diagnostic procedures  | <ul style="list-style-type: none"> <li>• Ultrasound-guided aspiration of fluid from organs and cysts/lesions</li> <li>• Tissue sampling with needles (biopsy)</li> <li>• Staging of internal cancer by endoscopic US (esophagus, prostate, rectum)</li> <li>• Evaluation of liver stiffness (fibrosis, cirrhosis)</li> <li>• Pregnancy (fetal development, placenta, umbilical cord flow)</li> </ul>   |
| Therapeutic procedures | <ul style="list-style-type: none"> <li>• Drainage of fluid collections (abscesses, cysts) by needle or catheter</li> <li>• Injection of drugs/electrodes into cancer masses or cysts</li> <li>• Breaking of urinary stones (Extra-corporeal shock wave lithotripsy)</li> <li>• Pain relief (carpal tunnel syndrome, chronic low back pain, shoulder pain)</li> <li>• High intensity focused ultrasound</li> <li>• Drug delivery</li> <li>• Pregnancy-related procedures</li> </ul> |

In 2011, Lau et al. patented a method for delivering high intensity focused ultrasound (HIFU) energy to a treatment site internal to a patient's body with the purpose of providing therapeutic treatment of internal pathological conditions, such as cancer [8,9]. At focal intensities 4-5 orders of magnitude

greater than diagnostic ultrasound (typically about 0.1 W/cm<sup>2</sup>), HIFU (typically about 1000-10,000 W/cm<sup>2</sup>) can induce lesions or tissue necrosis at a small location deep in tissue while leaving tissue between the ultrasound source and focus unharmed. Tissue necrosis is a result of focal temperatures typically exceeding 70° C.

Special transducers designed for high power ultrasound application are employed. Self focusing piezoceramic bowls made of low loss PZT (Lead Zirconate Titanate) or piezocomposite are widely used. A more expensive and technically more complex alternative is the use of phased array transducers composed of many single elements.

Another promising application of HIFU is drug delivery: in order to locally enhance the drug concentration in vivo in tumors, a drug can be administered either encapsulated in liposomic or other carrier bubbles intravenously. Once the tumor is sonicated, the bubbles are destroyed and the drug is released locally to the tumor. The resulting scenario is a local enhancement of the drug concentration in the focal area of the ultrasound beam. A wide spectrum of research has been conducted demonstrating proof of the concepts of the feasibility and effectiveness of this approach in vitro and in animal studies in vivo [10].

### III. ADVANCES OF ULTRASONIC METHODS APPLIED TO MATERIALS AND STRUCTURES

The directly measured quantities, ultrasonic velocity and attenuation, are required for the ultrasonic non-destructive technique of material characterization.

The ultrasonic velocity is related to density and elastic constants, such as tensile modulus, shear modulus, flexural modulus, bulk modulus, Young's modulus, Poisson's ratio, Lamè constants. The knowledge of elastic properties is basic to understand and predicting the behavior of engineering materials.

Based on velocity and attenuation measurement microstructure and morphology (such as mean grain size, grain size distribution, texture, anisotropy, density variations, etc.) and diffuse discontinuity (such as microcracking, microporosity, fiber breakage, impact damage, creep damage etc.) can be characterized. Ultrasonic assessments of mechanical properties (such as tensile strength, shear strength, yield strength, hardness, fracture toughness, fatigue resistance etc.) are indirect and depend on empirical correlations.

Ultrasonic tests are currently employed for advanced structural ceramics, for evaluating bond performance in adhesive bonding and for composites laminates. See, for example, [11-13]

Nowadays, ultrasonic testing also provides the characterization of advanced and smart materials, the synthesis and characterization of nanomaterials and nanofluids..See, for example, [15].

Nondestructive tests on bridges and buildings are carried out periodically for maintenance, performance, degradation

and quality assurance inspection. Three typical materials are used in these structures: steel, concrete and wood.

The typical structure of bridge consists of a substructure, a superstructure and a deck. The substructure is built with concrete, stone masonry, steel and wood. The superstructure is built with RC beams or rolled steel beams for short span, steel plate girders for intermediate span and steel trusses or arches for long span. The deck consists of steel plates and beams, prestressed concrete beams and timber.

In ultrasonic testing on steel structures [16] two mechanisms are detected: cracking and corrosion. Corrosion is detected with longitudinal waves and cracking is detected with shear waves. High frequencies are used in ultrasonic tests on steel structures, ranging from 2 to 10MHz.

In ultrasonic tests on wooden structures (see, for example, [17]) the damage due to aging, to fungi and borers attack and to mechanical actions are detected. The wave propagation in these structures is influenced by anisotropy. Frequencies used are the same as those for concrete ranging from 50 to 150kHz.

In ultrasonic tests on concrete structures [16], loss of strength due to many causes such improper mixture, chemical attack, microcracking, corrosion of steel rebar, fire damage can be detected by the traditional method called Pulse Velocity Measurement. The longitudinal wave pulse velocity is proportional to the square root of the elastic modulus and it is assumed that this modulus is proportional to square root of the compressive strength. Therefore there is an indirect fourth power relation between Ultrasonic Pulse Velocity and the compressive strength. The higher the velocity, the better the quality of the concrete. Crack depth measurement and thickness measurement can be also performed.

Another application of ultrasonics in testing concrete is the detection of rebar corrosion. Corrosion of RC structures has become a big problem worldwide due very high repair costs. Recently, the ultrasonic guided wave (UGW) technique is adopted to monitor the RC corrosion damage evolution process [18]. The corrosion experiment shows that the first wave peak value can describe the whole process of steel rebar corrosion.

#### *A. Application of Ultrasonic Pulse Velocity (UPV) test to concrete structures*

The UPV test is performed by using a sending transducer that sends an ultrasonic pulse to generate a stress wave in a concrete specimen and uses a receiving transducer to receive the wave.

By knowing the distance and time, the UPV in the specimen can be calculated. As pointed out by many researchers, the value of the UPV is affected by numerous factors, including the properties and proportion of the constituent materials, aggregate content and types, age of the concrete, presence of microcracks, water content, stresses in the concrete specimen, surface condition, temperature of the concrete, path length, shape and size of the specimen, presence of reinforcement, and so on.

The test equipment shown in Fig.2 consists of a pulse generator, a pair of transducers (transmitter and receiver), an amplifier, a time measuring circuit, a time display unit, connecting cables and material of coupling.

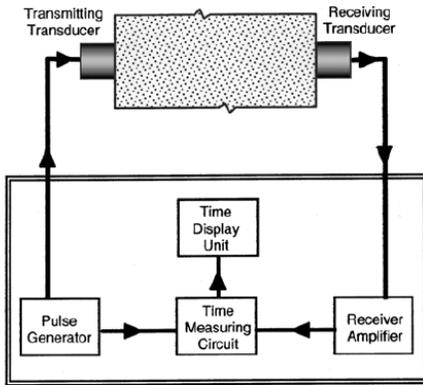


Fig.2 Schematic of Pulse Velocity Apparatus (ASTM C597)

The test is performed generating a series of pulses by means of a dedicated generator. A timing circuit is used to measure the travel time of the pulse, which is subsequently reported on the display unit. The presence of low density, or cracked, concrete increases the travel time and consequently causes a decrease in the pulse velocity. Performing tests at various locations, lower quality concrete or damage zone can be detected.

It is possible to make measurements of pulse velocity by placing the two transducers on opposite faces (direct transmission), on adjacent faces (semi-direct transmission), on the same face (indirect or surface transmission) of a concrete structure or specimen (Fig.3).

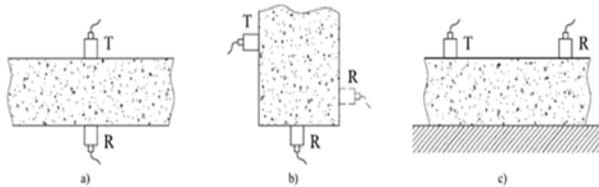


Fig.3 Schematics of transducers positioning: a) direct transmission; b) semi-direct transmission; and c) indirect or surface transmission.

In order to provide a correct measurement of ultrasonic pulse velocity, it is necessary to consider different factors which are able to influence pulse velocity and its correlation with mechanical and deformability characteristics of concrete. It can be summarized that:

- The moisture content has a strong effect on the pulse velocity, both for chemical and physical aspects.
- The path length over which the pulse velocity is measured should be long enough not to be significantly influenced by the heterogeneous nature of the concrete

- The presence of steel reinforcement can alter the ultrasonic pulse velocity.
- The presence of cracks and voids can considerably reduce the ultrasonic pulse velocity.

Even if a direct unique relationship between concrete compressive strength and ultrasonic pulse velocity does not exist, under specified conditions some information on its quality can be derived, as shown in Table 3.

Table 3. Quality of concrete as a function of the ultrasonic pulse velocity

| Ultrasonic Pulse velocity [m/s] | Quality of Concrete |
|---------------------------------|---------------------|
| > 4500                          | Excellent           |
| 3500 to 4500                    | Good                |
| 3000 to 3500                    | Doubtful            |
| 2000 to 3000                    | Poor                |
| < 2000                          | Very poor           |

According to the scientific literature in the field of ultrasonic investigations, there are numerous experimental formulations able to correlate the concrete compressive strength  $f_c$  with the measurement of the non-destructive parameter of the ultrasonic pulse velocity. Therefore, to limit the uncertainty of this method, three of the most reliable correlations proposed by Qasrawi [19] (Eq.A in Fig.3), Giannini et al. [20] (Eq.B in Fig.3), and Bilgehan and Turgut [21] (Eq.C in Fig.3) can be taken into account.

In order to evaluate the accuracy of these formulations, UPVs have been calculated in 16 locations by UPV test, as reported in [22],[23].

Then the estimated compressive strengths according to Eq. A, Eq. B and Eq. C, respectively, have been compared with the effective compressive strengths determined by Destructive Tests (DTs) on samples extracted in adjacent locations. To compare prediction performance of the formulations, Root Mean Square Error (RMSE) has been calculated. All the comparisons are shown in Fig. 3.

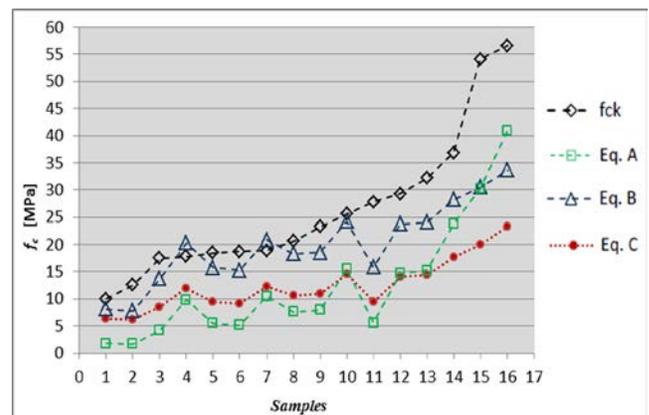


Fig.3 Comparison between DT  $f_c$  - values and estimated NDT  $f_c$

values using different formulations

By comparing all the correlation curves and the effective strength curve ( $f_c$ ), it can be observed a good approximation even if the most reliable correlation is related to Eq.B

#### IV. CONCLUSION

Ultrasonic methods represent a form of nondestructive testing and characterization of materials and structures used in engineering, including aerospace, automotive and other transportation sectors. The advantages of these methods include flexibility, low cost, in-line operation, and providing data in both signal and image formats for further analysis.

To explore the possibilities of the application of these emerging inspection technologies, the topic of the evaluation of structural performance of existing RC structures is considered in this work. Since these structures were built according to the standards and materials which were quite different to those available today, procedures and methods able to cover lack of data about mechanical material properties and reinforcement detailing are required. This issue seems more relevant when seismic zones are concerned and structural strengthening needs to prevent failures occurred due to earthquakes. Recent seismic codes give relevance to procedure and methods to establish the performance levels of existing structures. To this end detailed inspections and tests on materials are required. In RC structures, the compressive strength of concrete has a crucial role on the seismic performance and is usually difficult and expensive to estimate. According to various international codes, estimation of the in-situ strength has to be mainly based on cores drilled from the structure.

However, Non-Destructive Tests (NDTs) can effectively supplement coring thus permitting less expensive and more representative evaluation of the concrete properties throughout the whole structure under examination. The critical step is to establish reliable relationships between NDT results and actual concrete strength. The experimental research suggests correlating safely in most codes the results of in-situ NDTs carried out at selected locations with the strength of corresponding cores. As a consequence, NDTs can strongly reduce the total amount of coring needed to evaluate the concrete strength in the entire structure.

#### ACKNOWLEDGMENT

This research has been supported by the University of Bologna, Italy.

#### REFERENCES

- [1] F.A. Firestone, "Flaw detecting device and measuring instrument". US Patent 2280226,1942.
- [2] J. Ashman, E. Holm and B.Olstad, "Method for generating anatomical M-mode displays". US Patent 5515856 A,1996.
- [3] E.G. Tickner and N.S. Razor, "Ultrasonic image enhancement", US Patent 4276885 A,1981.

- [4] Y.Takeuchi, "Ultrasonic pulse doppler apparatus". US Patent 4759373 A,1988.
- [5] B.A.J. Angelsen and T.F. Johansen, "Extended, ultrasound real time 3D image probe for insertion into the body". US Patent 7699782 B2,2010.
- [6] B.R. Matlaga and J.E. Lingeman, "Surgical management of upper urinary tract calculi". In: Wein AJ, ed. Campbell-Walsh Urology. 10th ed. Philadelphia, Pa: Saunders Elsevier; chap 48,2011.
- [7] M.L. Palmeri, K.R. Nightingale, G.E. Trahey, K.D. Frinkley, "Methods, Systems and Computer Program Products for Ultrasound Shear Wave Velocity Estimation and Shear Modulus Reconstruction". US Patent 20080249408 A1,2008.
- [8] M. Lau, S. Vaezy, A.Lebedev and M.J.Connolly, "Apparatus for delivering high intensity focused ultrasound energy to a treatment site internal to a patient's body". US Patent 8057391 B2,2011.
- [9] M.Lau, N. Teng, S. Vaezy, A. Lebedev, M. W. Lau and M. J. Connolly "Methods and apparatus for the treatment of menometrorrhagia, endometrial pathology, and cervical neoplasia using high intensity focused ultrasound energy". US Patent 8277379 B2,2012.
- [10] O. Al-Bataineh, J. Jenne and P.O. Huber, "Clinical and future applications of high intensity focused ultrasound in cancer". *Cancer Treatment Reviews*, Vol.38, Number 5,2012,pp.346-353.
- [11] ASTM C1331 - 01. Standard "Test Method for Measuring Ultrasonic Velocity in Advanced Ceramics with Broadband Pulse-Echo Cross-Correlation Method",2012.
- [12] ASTM C1332 - 01. Standard "Test Method for Measurement of Ultrasonic Attenuation Coefficients of Advanced Ceramics by Pulse-Echo Contact Technique,2013.
- [13] G.B. Chapman, *Nondestructive Evaluation of Adhesive Bonds Using 20 MHz and 25 kHz Ultrasonic Frequencies on Metal and Polymer Assemblies*. Author House,2014.
- [14] Z. Zhoua, B. Maa, J. Jiang, G. Yu, K. Liu, D. Zhangb and W. Liu, "Application of wavelet filtering and Barker-coded pulse compression hybrid method to air-coupled ultrasonic testing. *Nondestructive Testing and Evaluation* Vol.29,Number 4, 2014;29: 297-314.
- [15] S.H. Sonawane, B.A. Bhanvase, R.D. Kulkarni and P.K. Khanna, *Ultrasonic processing for synthesis of nanocomposite via in situ emulsion polymerization and their applications. In Cavitation-A Novel Energy-Efficient Technique for the Generation of Nanomaterials*. Eds. S.Manickam and M.Ashokkumar. Pan Stanford Publishing,2014.
- [16] M.A.El-Reedy. *Concrete and Steel Construction: Quality Control and Assurance*. Taylor Francis Group, CRC Press,2013.
- [17] M. Krause, U. Dackermann and J. Li, "Elastic wave modes for the assessment of structural timber: ultrasonic echo for building elements and guided waves for pole and pile structures". *Journal of Civil Structural Health Monitoring*. DOI 10.1007/s13349-014-0087-2 F,2014.
- [18] D. Li, S. Zhang, W. Yang, W. Zhang, "Corrosion Monitoring and Evaluation of Reinforced Concrete Structures Utilizing the Ultrasonic Guided Wave Technique", *International Journal of Distributed Sensor Networks*. Article ID 827130,2014.
- [19] Y.H. Qasrawi, "Concrete strength by combined nondestructive methods simply and reliably predicted. *Cem Concr Res*, Vol.30, 2000, pp.739-746.
- [20] R. Giannini, L. Sguerri and V. Ninni, "Affidabilità dei metodi d'indagine non distruttiva per la valutazione della resistenza del calcestruzzo (in Italian)". *Proceedings of 10° Congresso Nazionale AIPnD*, Ravenna, Italy,2003.
- [21] M.P. Bilgehanand P. Turgut, "Artificial Neural Network Approach to Predict Compressive Strength of Concrete through Ultrasonic Pulse Velocity", *Res Nondestr Eval*, Vol.21, 2010,pp. 1-17.
- [22] L. Nobile, "Prediction of concrete compressive strength by combined non-destructive methods". *Meccanica*, Vol. 50, Number 2 (2015) pp.411-417.
- [23] L. Nobile, and M. Bonagura, "Recent advances on non-destructive evaluation of concrete compression strength". *Int. J. Microstructure and Materials Properties*, Vol. 9, Numbers 3-5/2014,pp.423-421.

# Mobility State Classification with Particle Filter

Ha Yoon Song, Ji Hyun Baik

**Abstract**—Positioning data can be obtained using various mobile devices nowadays. The consequent positioning data can be used to figure out speed values with latitude, longitude, and timestamp. However, the speed value itself cannot be criteria, whether the corresponding device is moving or staying due to positioning system errors. For the sophisticated classification of mobility state, an algorithm based on particle filter is introduced. The speed values are distilled by the algorithm using particle filter and the mobility state can be probabilistically identified. On the basis of previous research, a set of consecutive speed values is used for particle filter in order to enhance the preciseness. The algorithm and related experimental results will be presented.

**Keywords**—Human Mobile State Determination, Particle Filter, Positioning Data, Probabilistic Approach.

## I. INTRODUCTION

POSITIONING data can be easily obtained nowadays due to advances of mobile devices such as smartphones. Of course, these positioning data can be directly used for other applications, however, there are various reasons for the positioning data to have errors. Therefore the positioning data must be filtered or classified in case of errors, i.e. classification of mobility states must be included priory for the use of positioning data. These sorts of classification cannot be deterministically made, however it is possible to probabilistically make such classification. In this paper, the simplest classification of human mobility state will be addressed by the proposed algorithm. The simplest classification is to divide human mobility state in *mobile* or *stable* state probabilistically. This sort of classification would add more accuracy to positioning data related applications. Section II will describe related works. In section III, an algorithm for classification of mobility states using particle filter will be addressed. In section IV we will show the results of algorithm application in various ways. Section V will discuss on conclusion and possible future research.

## II. RELATED RESEARCH

### A. Particle Filter

Particle Filter is so called Sequential Monte Carlo (SMC) and based on Bayesian statistics and is usually used for parameter estimation, state estimation and so on. The basic idea of particle filter is to generate and utilize a large number of independent random variables. The independent random variables are called particles. The values of particles are set

by initialization through state space. The values of particles are updated by weights as inputs which is newly measured observations [1]. There are two representative particle filter algorithms, Sampling Importance Resampling (SIR) and Sequential Importance Sampling (SIS). SIS does not include resampling process, whereas SIR requires resampling process. We used SIS which is monte carlo method, constructing Sequential Monte Carlo Filter [2].

The action of SIS based algorithm likes the following:

- 1) Initialize particles  $X$ .
- 2) Obtain new observations  $Z$  and update likelihood probability  $P(Z|X)$ .
- 3) Obtain weights  $W$  using likelihood probability and update values of particles  $X$ .
- 4) Repeat steps 2 and 3 to update all particles.

### B. Previous Research with Particle Filter

In the field of computer science, particle filters are widely adopted for the position estimation of human with Wi-Fi signal, the position estimation of Robot and so on.

1) *Position Estimation with Wi-Fi signal and Wi-Fi terminal*: This is a sort of indoor positioning technique. The spaces are represented based on graphs, and the position of terminal holder can be represented as a point on edge of the graph using the signal strength of Wi-Fi system. The reason why the particle filter is introduced is that it is easy to implement and it is accurate [3].

2) *Particle Filter for Robot Position Estimation*: In continuous state space, particle filter is usually used for robot related area. Especially for mobile robot, it is quite successful where the relative position presented on the map through the sensors and actuators by coordinates and directions. The major benefit of particle filter is that it can be used in higher dimensional space [4].

### C. Basic Application of Particle Filter to Mobility State Classification

In our previous research [5] the very basic approach was made in order to determine the mobility states. Only the position data was utilized in previous researches, while the weight sequence of speed values is introduced and the distribution of human speed values as exponential distribution is utilized in this research. However there were several problems. The basic approach which directly uses particle filter for speed values is somewhat inaccurate, i.e. slow speed particles are regarded as stay state which is not true. And the toggle of state between mobile and stay is usually happened due to positioning errors. Thus it is considered to utilize history of speed values. From the next section, we will show the enhancement of our previous application of particle filter. The

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2012R1A2A2A03046473).

Ha Yoon Song is with the Department of Computer Engineering, Hongik University, Seoul, Korea (e-mail: hayoon@hongik.ac.kr).

Ji Hyun Baik is with the Department of Computer Engineering, Hongik University, Seoul, Korea (e-mail: jihyun135790@gmail.com).

major differences between previous research and this research are that Time Weighted Speed is introduced as shown in section III-A and the validation was executed with more than 3 million positioning data to encompass the preciseness of our algorithm.

---

**Algorithm 1** Mobility State Classification with Particle Filter
 

---

```

1: procedure
2: CLASSIFICATION( $N, V_{t-1}, V_{t-2}, V_{t-3}, V_{t-4}$ )
3:   Obtain new position
4:   Calculate  $V_t$ 
5:     ▷ Calculate current speed by haversine formula
6:   Calculate  $T$ 
7:     ▷ Calculate time weighted speeds by (1)
8:    $Particle[N]$      ▷ Declare and initialize Particles
9:   for  $i = 0$  to 4 do
10:    for  $k = 0$  to  $N - 1$  do     ▷ For  $N$  Particles
11:       $W = f(Particle[k], T)$      ▷ Weight update
12:    if condition then
13:       $Particle[k] = Particle[k] - W$ 
14:    else
15:       $Particle[k] = Particle[k] + W$ 
16:    end if     ▷ Particle update
17:  end for
18: end for
19:  $X = Average(Particle[N])$      ▷ Calculate  $X$ 
20: return  $X$ 
21:     ▷ Mobility State Classification
22: end procedure
    
```

---

### III. ALGORITHM

#### A. Time Weighted Speeds

In order to solve problems occurred in past researches, Time Weights Speed, which is giving weights to the appropriate number of past speed values, is introduced in this research.

Positioning data errors lead to the speed value errors and thus lead to mobility state errors. Thus the speed values in past history are also utilized in this research. The past speed values are weighted by its timestamp. The more recent the timestamp is, the higher the effect of speed is. It is presumably true that recent speed values are highly related to current speed values than other older speed values. An exception occurs only when mobile state changes from stay to mobile or mobile to stay. There exist less relationship between past and current speed values. Therefore, several consecutive speed values compose one window instead of using only two consequent speed values. The number of speed values to construct speed value window is called window size, and it can be set as an arbitrary natural number. The speed values in this window will be utilized to calculate Time Weighted Speeds.

The weighted average of time weighted speeds,  $T$ , is introduced and utilized. The value  $T$  is utilized to obtain the weights of particles. The rate which determines the weight of speed values, denoted as  $\alpha$ , is introduced. Of course the value of  $\alpha$  is between 0 and 1. We used total five speed

values, including one current speed value and four past values to calculate  $T$ .

With current speed  $V_t$  and past four speeds  $V_{t-1}, V_{t-2}, V_{t-3}, V_{t-4}$  where current time is  $t$ .  $V_{t-1}$  is a more recent speed value than  $V_{t-4}$ .  $T$  can be calculated as shown in (1) in case of window size as five. Once  $\alpha$  is set as 0.0, only the current speed will be used for particle filter, whereas 1.0 as  $\alpha$  stands that only the less recent speed value  $V_{t-4}$  will be used for particle filter.

$$T = (1 - \alpha)V_t + \alpha((1 - \alpha)V_{t-1} + \alpha((1 - \alpha)V_{t-2} + \alpha((1 - \alpha)V_{t-3} + \alpha(V_{t-4})))) \quad (1)$$

Window size of five is found to be optimal from experiments. More than four times of multiplications of  $\alpha$  leads to very small positive value close to 0 since  $\alpha$  is smaller than 1, and then can be disregarded with very negligible effects on  $T$ . Thus, window size is determined as five after simple experiments which checks the effects of window size with window sizes up to nine. With variations of windows size from 5 to 9, there are no clear differences in results.

#### B. Input Data Preparation

1) *Positioning Data Collection*: Numerous positioning data have been collected from November 2011. Most of data have been collected using smartphone app such as Sports Tracker [6]. As well, dedicated GPS receivers such as Garmin Edge 810 or similar devices [7] have been used. Among the various fields of positioning data set, latitude, longitude and time information are extracted and arranged as input sets of this research.

2) *Calculation of Speed between Positions*: With two consecutive positioning data, it is possible to calculate the distance between two points. Among the several methods to calculate the distance, Haversine formula [8] is used because of its simplicity. The only exception is the very first positioning data. With the time information embedded in positioning data, the speed values can be calculated. Then it is possible to have consecutive speed values.

#### C. Algorithm for Mobility State Classification with Particle Filter

Algorithm 1 shows the outline of particle filter based classification. Here, SIS particle filter methodology is used. It needs to reserve four past speed values as well as obtain the current positioning data. For every input data, it classifies mobility state, i.e. the algorithm must be called whenever new positioning data are acquired.

The overall action of algorithm 1 follows:

- 1) For new input of positioning data, calculate the speed value ( $Z$ ) at the given time by Haversine formula (line 4).
- 2) Calculate time weighted speeds ( $T$ ) using (1) (line 6).
- 3) Create  $N$  particles and initialize them (line 8).
- 4) Using (2) and (3), calculate weight  $W$  (line 11).
- 5) Update probability of particle (line 12-16).

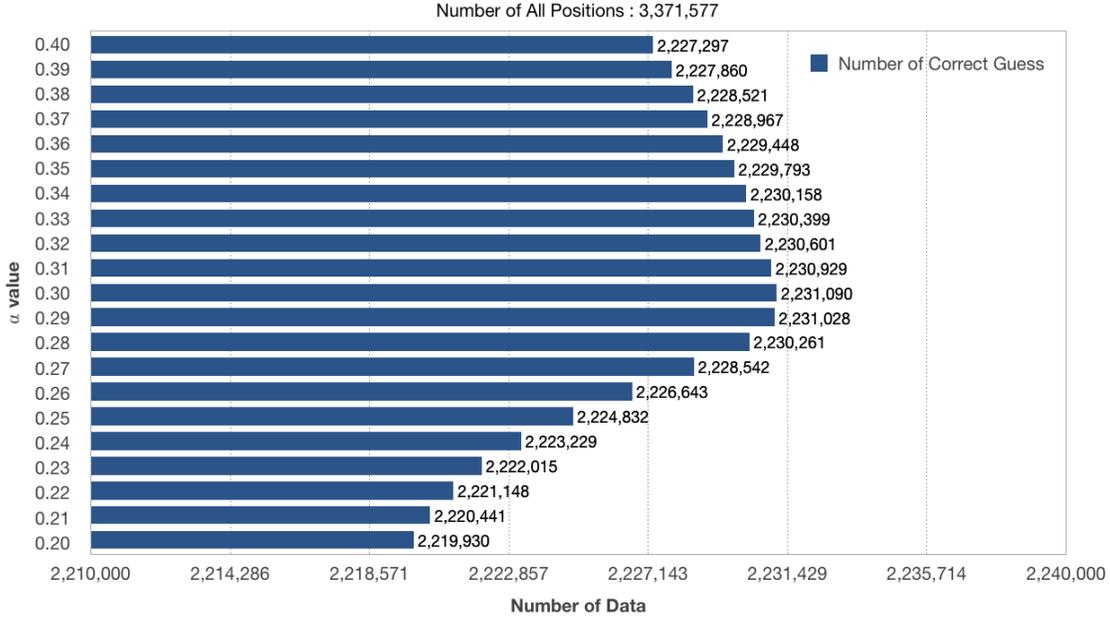


Fig. 1: Number of Correct Classifications

- 6) Repeat steps 4 and 5 to update  $N$  particles (line 10).
- 7) Repeat inner for loop (line 10) 5 times (line 9).
- 8) Calculate the average of particles (line 19), and determine the mobility state (line 20).

$$P_t = Pr[T = t] = 1 - e^{-\lambda T} \quad (2)$$

$$f(\text{Particle}[k], T) = \text{Particle}[k] - P_t/C \quad (3)$$

Then, the details of algorithm 1 need to be explained.

The algorithm has five inputs. The inputs are four past speed values as well as the number of particles,  $N$ . With the new positioning data, the current speed value is calculated. This five speed values compose a window of speed values and window size is five, except the very first stage of classification where there are less than five speed values. After generating  $N$  particles, initialize the probability of particles randomly, but having 0.5 as average. The purpose of randomization is to prepare so that particles can represent every possible state, locations in this research [4].

It is reported that various distributions can be used to represent the probability distribution of human speed and the exponential distribution is the simplest one [9]. Therefore, in order to represent the probability for  $T$ , exponential distribution is adapted.  $P_t$  is a probability such that  $T$  has a certain speed value  $t$ . Based on exponential distribution, cumulative distribution function of exponential distribution is introduced as shown in (2).

In algorithm 1,  $f(\text{Particle}[k], T)$  is used to obtain the weight as expressed in (3).  $P_t$  is calculated by (2) and then used in (3). Here, parameter  $\lambda$  is referred from [9] as 0.15949. The value of  $\lambda$  as 0.15949 stands that the expectation and standard deviation of speed distribution is 6.27m/s. In (3), constant  $C$  is introduced to have a normalized value of  $P_t$ , and is 4.0 in this research.

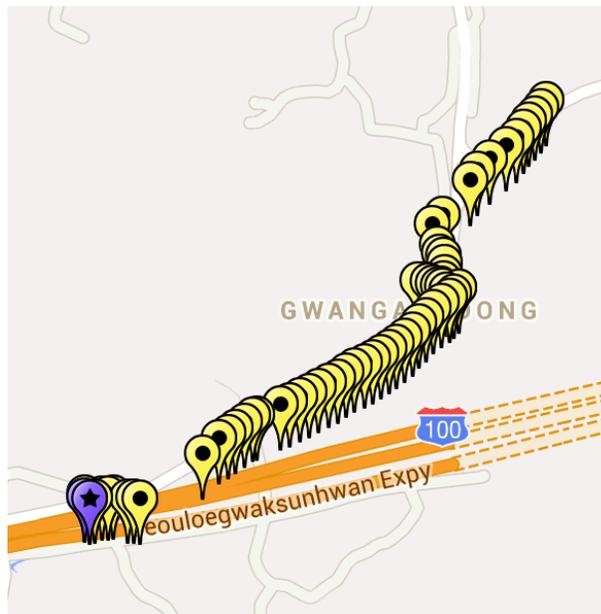
The constant 5 in algorithm 1 is used to repeat the update of  $N$  particles in order to stabilize the probability value and the accumulated value of particles in 5 times are utilized, on the contrary that usual SIS algorithms update the particle values only one time.

| Condition         | Particle $\leq 0.5$ | Particle $> 0.5$ |
|-------------------|---------------------|------------------|
| Weight $\leq 0.5$ | line 15 (+)         | line 13 (-)      |
| Weight $> 0.5$    | line 13 (-)         | line 15 (+)      |

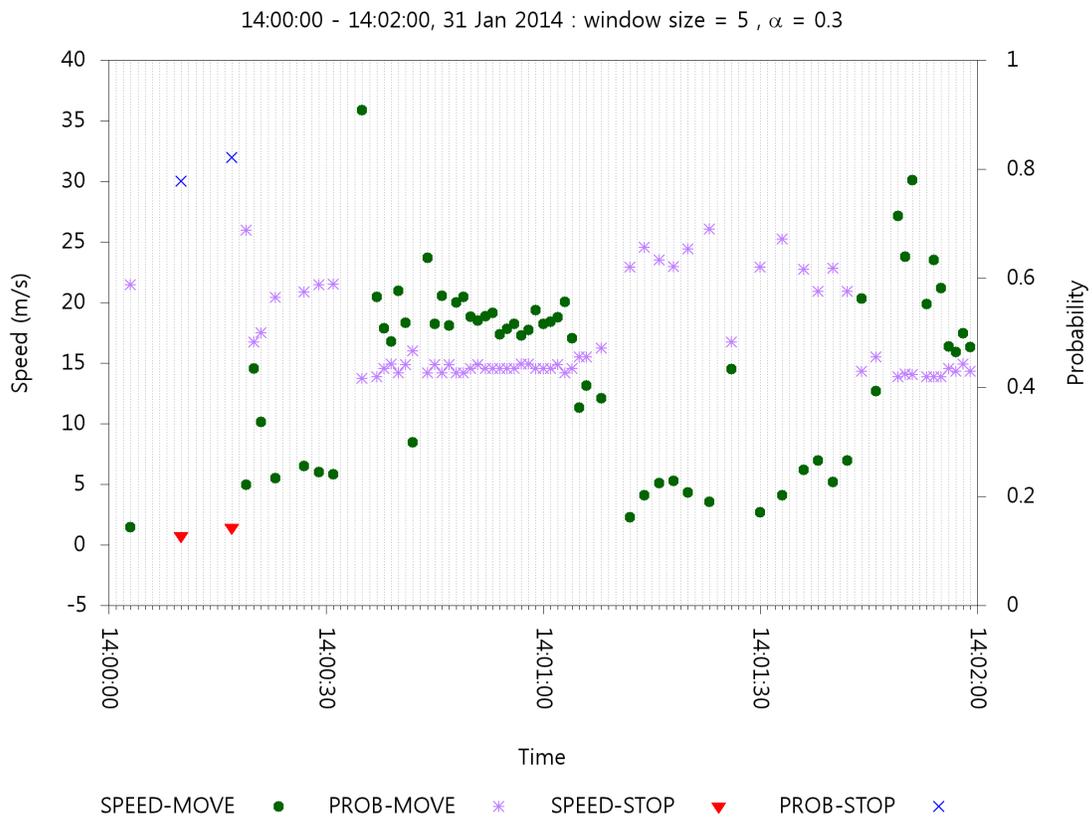
TABLE I: Particle update conditions

Algorithm 1 imposes a condition to update particle as shown in (line 13) and (line 15). This is a kind of amplification to put particle values closer to 0 or 1. The condition can be found in Table I. Expression in (line 13) shows particle value - weights, and expression in (line 15) has particle value + weights. When weight is big enough, the small particle value will be smaller, while the big particle value will be bigger. Consequently, weight will be subtracted to make small particle value smaller while weight will be added to make big particle value bigger. On the other hand, with small weight, big particle became small and small particle value became big. Consequently, weight will be added to make small particle value bigger while weight will be subtracted to make big particle value bigger. The criterion particle value and weight to be big or small is 0.5 in this algorithm.

Some miscellaneous adjustments have been made to cope with negative probability values which are abnormal, then such negative probability values are set to zero. Or the probability value is also set to one where the probability is greater than one. Finally, with the stabilized average probability  $X$ , the decision of mobility states can be made. In our implementation, 0.75 is the empirical threshold of state classification for state classification. The constant for state classification and constant for repetition must be determined by users of this algorithm.



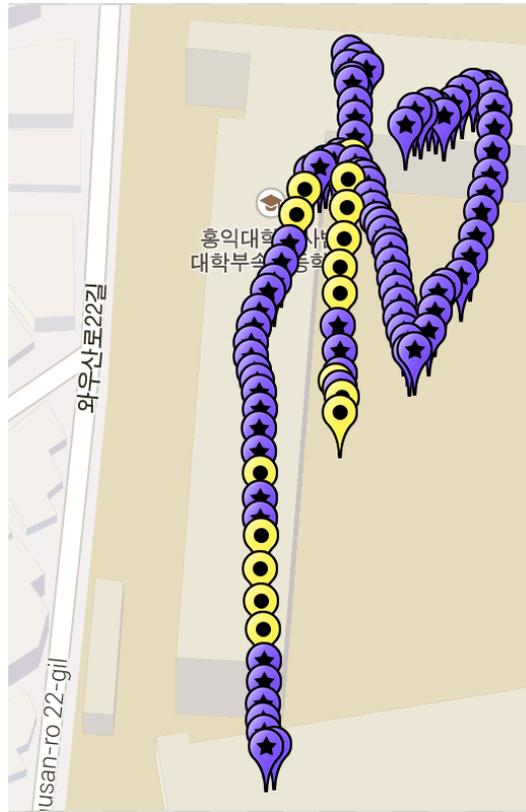
(a) Mobility State on Map



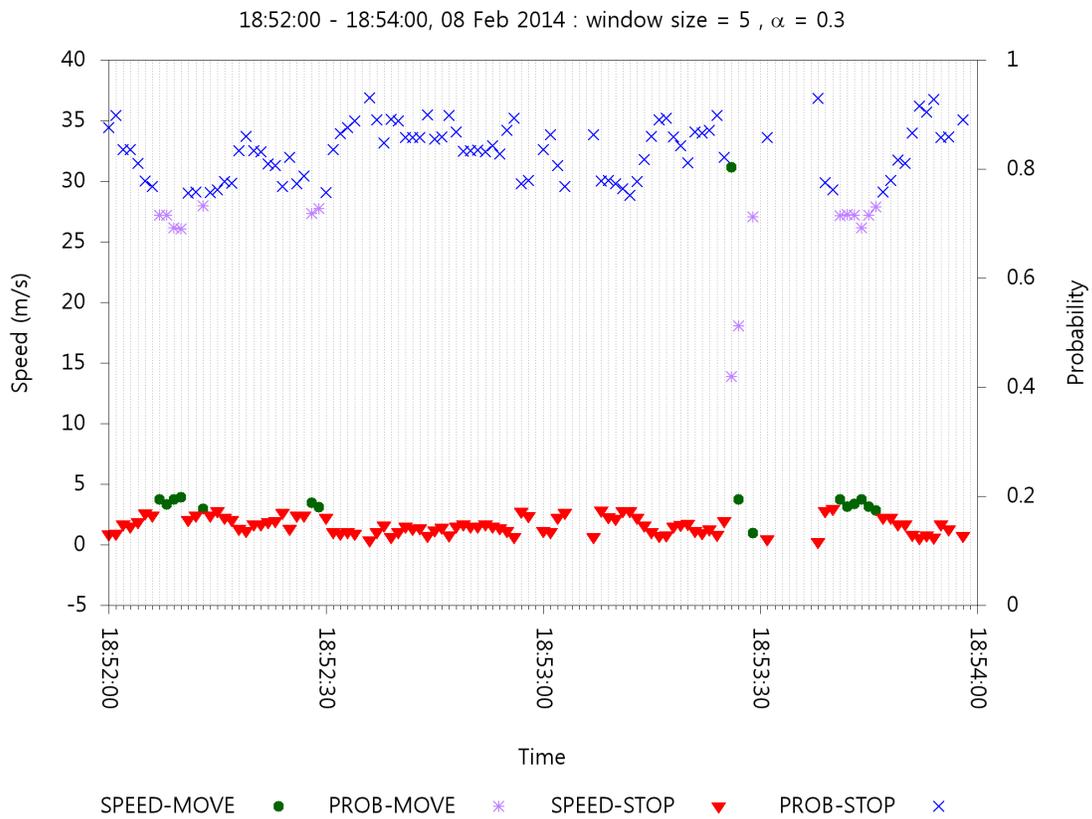
(b) Speeds-Probability Graph

Fig. 2: Mobility State Classification on 31 Jan 2014

The particles with a probability value greater than 0.75 are classified as stay, and otherwise as mobile.

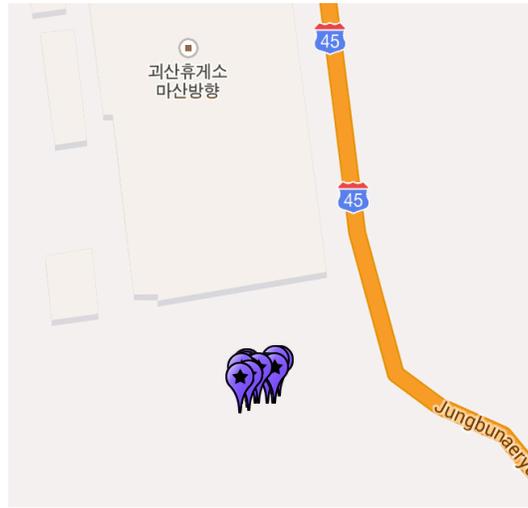


(a) Mobility State on Map

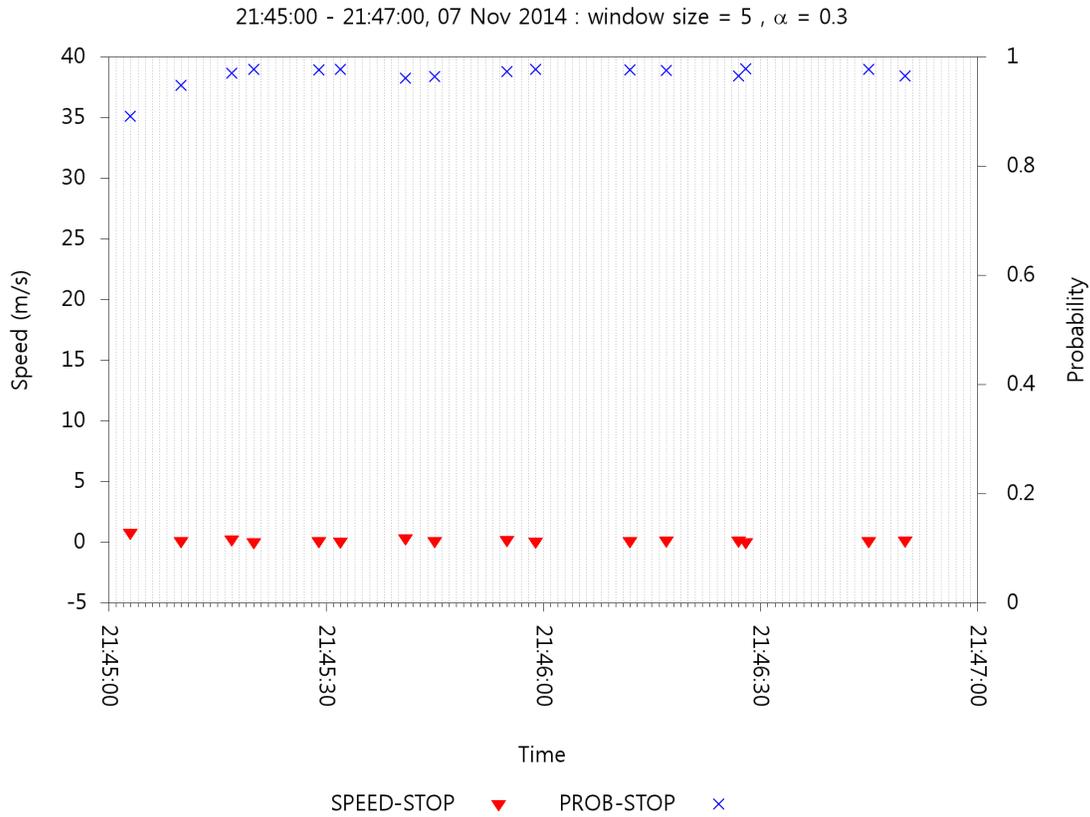


(b) Speeds-Probability Graph

Fig. 3: Mobility State Classification on 08 Feb 2014



(a) Mobility State on Map



(b) Speeds-Probability Graph

Fig. 4: Mobility State Classification on 07 Nov 2014

#### IV. APPLICATION OF ALGORITHM: RESULTS AND EXPLANATION

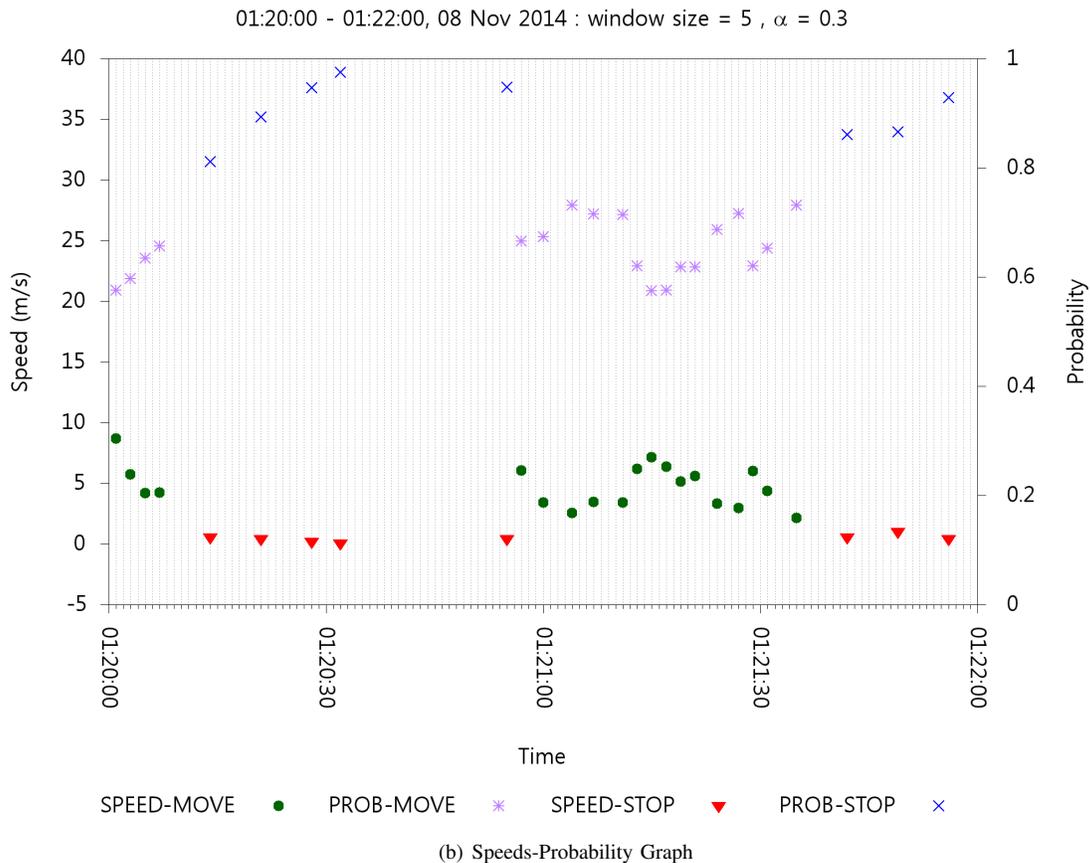
##### A. Searching for optimal weight parameter $\alpha$

In order to implement time weighted speeds in (1), constant  $\alpha$  need to be determined. It is usually required to find the optimal  $\alpha$ . Fortunately, the region of  $\alpha$  is  $(0,1)$  which is very restricted to apply exhaustive search. With 3,371,577 positioning data,  $\alpha = 0.3$  shows the optimal result where

2,231,090 positioning data shows exact classification compared with manual classification. As a result the percentile of precision is 66.173 when  $\alpha$  is 0.3. Fig. 1 shows the correct classification with various values of  $\alpha$  from 0.20 to 0.40, where X-axis is the number of positioning data and Y-axis is the value of  $\alpha$  and the numbers by bar in the graph is the number of correct classifications. When  $\alpha = 0.0$ , 2,146,439 samples guessed correct, whereas 198,161 samples guessed



(a) Mobility State on Map



(b) Speeds-Probability Graph

Fig. 5: Mobility State Classification on 08 Nov 2014

correct when  $\alpha$  is 1.0. It implies most of guess are incorrect when  $\alpha$  is 1.0. As discussed in section III-A, the fourth past speed value has very little relationship to the current mobile state when  $\alpha$  is 1.0. Therefore, window size more than five would be meaningless.

**B. Result of Algorithm application**

In subsection IV-A, the optimal  $\alpha$  as 0.3 has been found. Fig. 2, Fig. 3, Fig. 4, and Fig. 5 show the classification result by algorithm 1. Each figure shows the collected data in two minutes projected on google map and corresponding graph where window size is five. Each figure contains speeds-probability graph. For each graph, the X-axis stands for time of

a day, left Y-axis stands speeds of positioning data in meter per second at given time, and right Y-axis stands for probability calculated by algorithm 1.

Placemarks are used to designate states on the map: stands for stay state and stands for mobile state. Every placemark corresponds to positioning data, respectively. For each graph, markers are also used to designate speed values and probabilities: stands for the speed value for mobile state position, stands for the average probability of particles for mobile state position, whereas X stands for the average probability of particles for stay state positions and stands for the speed value for stay state position.

Fig. 2 shows the case of automobile movement. Fig. 2(b)

shows that the most of probability values reside between 0.4 and 1.0. It is notable that the speed of automobile varies widely. It shows a typical example of positioning error around time 14:00:35 showing the speed value about 35m/s. Compared to the neighboring values, it can be regarded as positioning error. A speed value around time 14:01:10 is classified probabilistically as mobile, regardless of its value of 3m/s, because the neighboring speed values are also mobile.

Fig. 3 shows the case of ambulation in university campus. Most of positions are regarded as stay with low speed values. The speed values distribute between 0m/s and 5m/s, smaller variations can be found with relatively constant speed values comparing to the case of automobile as shown in Fig. 2(b). At time 18:53:25, a high speed value is observed abruptly. The conjecture of this abnormality is that positioning data error due to urban canyon phenomena.

Fig. 4 shows the case of expressway rest area. As expected, all of speed values are small enough and all classifications are determined as stay. Since all of speed values designate stay states as shown in Fig. 4(b), Fig. 4(a) shows similar positions on the map. The actual mobile state observed is stay, and classification is correct, even though the positioning system error, where the positioning system report fluctuating positions even though the mobile state is stay. It can be concluded that our algorithm by speed values with a sophisticated mechanism enhance the correctness of classification.

Fig. 5 shows the case of automobile ride, with clear variations in speed values. Some part of consecutive data show slow speed and high probability of stay, whereas other parts of data show high speed and high probability of move. It does stand for typical stop and go situation of automobile driving. In Fig. 5(b), around 01:20:13, five consequent values are classified as stay, which correspond to stay placemarks at the lower left part of Fig. 5(a). It is a U-turn of a vehicle and relative small speed values compared to other speed values lead to the classification of stay also with large time differences. And the vehicle starts acceleration neat 01:20:53.

## V. CONCLUSION AND FUTURE RESEARCH

Particle filter is used in order to classify mobile state with time weighted speeds values on the basis of preliminary research [5]. Proper weight parameter has been identified with 3,371,577 sample values. The value of optimal weight parameter  $\alpha$  is 0.30.

There is one point of future improvement. In cases of ambulation or traffic jam, low speed values are continuously observed and these values are probabilistically regarded as stay state. Another point of future research is calculating the parameter of exponential distribution  $\lambda$  according to the situation of given time. In detail, speed values in a window will be used to calculate scale parameter and location parameter of exponential distribution. It is expected that dynamic calculation of parameters will solve the problems aforementioned since these two problems are not separated ones. With parameters upon the situations, particle filter is expected to show more precise classification.

Another topic is to classify the mobile state to more categories; e. g. ambulation, bicycle, automotive, train, airplane

and so on since every of these transportation method has its own speed characteristics.

Finally, the classification of micro or narrow mobility, and macro or broad mobility will be able to be introduced. For example, mobility inside a specific building complex (micro mobility) and exit outside the specific area (macro mobility) can be classified. This topic is not only related with the probability distribution speeds but also related with the probability distribution of location or position.

## REFERENCES

- [1] Z. Chen, "Bayesian filtering: From kalman filters to particle filters, and beyond," *Statistics*, vol. 182, no. 1, 2003, pp. 1–69.
- [2] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, Feb 2002, pp. 174–188.
- [3] E. M. Choi, H. K. Oh, and I. C. Kim, "Particle filters for positioning wifi device users," *Journal of KIISE : Software and Applications*, vol. 39, no. 5, 2012, pp. 382–389.
- [4] S. Thrun, "Particle filters in robotics," in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 511–518. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2073876.2073937>
- [5] J. H. Baik and H. Y. Song, "Determining human movement with particle filter," *The 2014 Fall Conference of the KIPS 21(1) 2*, vol. 1, no. 1-12, 2014, pp. 372–375.
- [6] Sports Tracker. [Online]. Available: <http://www.sports-tracker.com>
- [7] Garmin. [Online]. Available: <http://www.garmin.com/en-US>
- [8] R. W. Sinnott, "Virtues of the haversine," *Sky and Telescope*, vol. 68, no. 2, 1984, p. 159.
- [9] H. Y. Song and J. S. Lee, "Finding probability distributions of human speeds," in *AMBIENT 2014, The Fourth International Conference on Ambient Computing, Applications, Services and Technologies*, 2014, pp. 51–55.

**Ha Yoon Song** Ha Yoon Song received his B.S. degree in Computer Science and Statistics in 1991 and received his M.S. degree in Computer Science in 1993, both from Seoul National University, Seoul, Korea. He received Ph.D. degree in Computer Science from University of California at Los Angeles, USA in 2001. From 2001 he has worked at Department of Computer Engineering, Hongik University, Seoul, Korea and is now a professor. In his sabbatical year 2009, he worked at Institute of Computer Technology, Vienna University of Technology, Austria as a visiting scholar. Prof. Song's research interests are in the areas of mobile computing, performance analysis, internet of things and human mobility modeling.

**Ji Hyun Baik** Ji Hyun Baik is a student at Hongik University, Seoul, Korea and majors in Computer Engineering. Her interested areas are mobile computing, big data analysis and database.

# The impact of memristive devices and systems on nonlinear circuit theory

Ricardo Riaza

*Abstract*—In this talk we present a discussion of the impact of memristive devices (memristors, memcapacitors and meminductors) and memristive systems on the fundamentals of nonlinear circuit theory and also on electronics. The interest on this topic has increased continuously since the design in 2008 of a nanoscale device with a charge-flux characteristic, displaying a very promising industrial scope in memory design. We survey theoretical concepts and also discuss some recent applications in this area, providing references for further study.

*Keywords*—Nonlinear circuit, memristor, memcapacitor, meminductor, nonlinear oscillator, chaotic circuit, resistive memory, memristive neural network.

## I. INTRODUCTION

**T**HE history of memristive devices can be traced back to the seminal work of Leon Chua, who in 1971 postulated the existence of a nonlinear device whose characteristic would be defined by a charge-flux relation [1]. This device would be the fourth basic circuit element, besides the resistor, the inductor and the capacitor which relate the voltage-current, current-flux and voltage-charge pairs, respectively. The report in 2008 of a nanometer-scale device displaying a memristive characteristic [2] has had a great impact in the electrical and electronic engineering communities and has raised a renewed interest towards these devices.

The memristor and related devices are likely to play a relevant role in electronics in the near future, especially at the nanometer scale. Many applications are already reported, e.g. in pattern recognition, memory design, signal processing, design of nonlinear oscillators and chaotic systems, adaptive systems, etc. (see [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]). Commercial memory chips based on the memristor are expected to be released in the near future. The notion of a device with memory was extended to the reactive setting by Di Ventra, Pershin and Chua [8] in order to define *memcapacitors* and *meminductors*.

We survey in this talk some basic properties of memristors, memcapacitors and meminductors (cf. Section II) together with those of the so-called *memristive systems*, introduced by Chua and Kang in [31] (Section III). The impact of such devices and systems on the fundamentals of nonlinear circuit theory and electronics is surveyed in Section IV, whereas Section V compiles some applications of these devices and systems in electronics. Concluding remarks are compiled in Section VI.

R. Riaza is with the Departamento de Matemática Aplicada a las TIC, ETSI Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain. [ricardo.riaza@upm.es](mailto:ricardo.riaza@upm.es)

## II. MEMRISTORS, MEMCAPACITORS AND MEMINDUCTORS

### A. Memristors

The memristor is a nonlinear device defined by a charge-flux characteristic, which may be have either a charge-controlled or a flux-controlled form. In a charge-controlled setting, the characteristic reads as

$$\varphi = \phi(q), \quad (1)$$

for some  $C^1$  map  $\phi$ . The incremental *memristance* is

$$M(q) = \phi'(q).$$

Using the relations  $\varphi' = v$ ,  $q' = i$  we get the voltage-current characteristic

$$v = M(q)i. \quad (2)$$

This relation shows that the device behaves as a resistor in which the resistance depends on  $q(t) = \int_{-\infty}^t i(\tau)d\tau$ , hence the *memory-resistor* name. This is the key feature of the device. In greater generality, one may consider (2) as a particular case of a fully nonlinear characteristic of the form

$$v = \eta(q, i).$$

In turn, a flux-controlled memristor has a characteristic of the form

$$q = \xi(\varphi), \quad (3)$$

and the incremental *memductance* is

$$W(\varphi) = \xi'(\varphi).$$

The voltage-current relation has in this case the form

$$i = W(\varphi)v \quad (4)$$

or, in a fully nonlinear context,

$$i = \zeta(\varphi, v).$$

A memristor governed by (2) or (4) is said to be *strictly locally passive* if  $M(q) > 0$  or  $W(\varphi) > 0$  for all  $q$  or  $\varphi$ , respectively. In the presence of coupling effects (if eventually displayed), this requirement must be restated by asking the memristance or memductance matrices to be positive definite.

### B. Memcapacitors and meminductors

Di Ventra, Pershin and Chua extended in [8] the idea of a device with memory to reactive elements. A (voltage-controlled) *memcapacitor* has a characteristic of the form

$$q = C_m(\varphi)v. \quad (5)$$

Here  $C_m$  is the *memcapacitance*. The distinct feature of this device is that the memcapacitance depends on the state variable  $\varphi(t) = \int_{-\infty}^t v(\tau)d\tau$ , so that the relation  $q(t) = C_m(\int_{-\infty}^t v(\tau)d\tau)v(t)$  reflects the device history. Analogously, a (current-controlled) *meminductor* is governed by

$$\varphi = L_m(q)i, \quad (6)$$

and  $L_m(q)$  is the *meminductance*, which reflects the device history via the variable  $q$ .

A fully nonlinear formalism for voltage-controlled memcapacitors and current-controlled meminductors is obtained after replacing (5) and (6) by the characteristics

$$q = \omega(\varphi, v) \quad (7)$$

and

$$\varphi = \theta(q, i), \quad (8)$$

respectively, for certain maps  $\omega, \theta$ .

### III. MEMRISTIVE SYSTEMS

Chua and Kang extended in [31] the ideas underlying the notion of a memristor to define a *memristive system*. A current-controlled memristive system is defined by a relation of the form

$$q' = \mu(q, i) \quad (9a)$$

$$v = M(q, i)i, \quad (9b)$$

whereas a voltage-controlled one is defined by

$$\varphi' = \psi(\varphi, v) \quad (10a)$$

$$i = W(\varphi, v)v. \quad (10b)$$

In particular, a charge-controlled memristor is an instance of (9) in which  $q$  represents charge,  $\mu(q, i)$  amounts to the current  $i$  and  $M$  does not depend on  $i$ , as discussed in Section II. In this case,  $M(q)$  is the (incremental) memristance, as indicated above; this arises as the derivative of a nonlinear flux-charge relation  $\varphi = \phi(q)$ , and this supports the “charge-controlled” term. Similarly, a flux-controlled memristor is a particular instance of (10) in which  $\varphi$  stands for the magnetic flux,  $\psi(\varphi, v)$  amounts to the voltage  $v$  and  $W$  does not depend on  $v$ . Now  $W(\varphi)$  is the (incremental) memductance.

In general, the key feature of memristive systems is that the memristance  $M(q, i)$  in (9b), or the memductance  $W(\varphi, v)$  in (10b), depend on a state variable (denoted as  $q$  and  $\varphi$ , respectively) which is governed by a differential equation (namely (9a) and (10a)). The values of  $q$  and  $\varphi$  keep track of the device history since they must be computed (or, more precisely, they are defined) as an integral variable. Note that, for the sake of notational simplicity, we use  $q$  and  $\varphi$  to represent the dynamic variables of general memristive devices,

although only for memristors in strict sense these variables denote charges and fluxes, respectively.

Apart from “classical” memristors, other instance of memristive devices are thermistors, discharge tubes and ionic systems. In all these systems (and all in “classical” memristors), the matrices  $M$  and  $W$  are actually independent of  $i$  and  $v$ , respectively, yielding the simpler forms

$$q' = \mu(q, i) \quad (11a)$$

$$v = M(q)i \quad (11b)$$

in the current-controlled context, and

$$\varphi' = \psi(\varphi, v) \quad (12a)$$

$$i = W(\varphi)v \quad (12b)$$

in a voltage-controlled setting.

Many recent applications of memristive devices are better framed in the context defined by (9) and (10) (or in the simplified contexts represented by (11) and (12)) rather than in the setting defined in Section II. This is also the case for memcapacitive and meminductive systems, as discussed in [8].

### IV. HIGHER ORDER DEVICES AND THE FUNDAMENTALS OF NONLINEAR CIRCUIT THEORY

Going back to the context defined by memristors, memcapacitors and meminductors (cf. Section II), it is worth noticing that both (5) and (6) come from differentiating a two-variable relation, namely  $\sigma = \alpha(\varphi)$  for voltage-controlled memcapacitors and  $\rho = \beta(q)$  for current-controlled meminductors; here  $\sigma$  and  $\rho$  arise as the time-integrals of  $q$  and  $\varphi$ , respectively. By using the differentiated relations (5) and (6) we get rid of these second order variables. This is not the case for so-called *second-order devices*, for which either  $\sigma$  or  $\rho$  appear explicitly in the memcapacitance or the meminductance. Specifically, a *charge-controlled memcapacitor* is a device defined by the relations

$$\sigma' = q \quad (13a)$$

$$q' = i \quad (13b)$$

$$v = C^{-1}(\sigma)q, \quad (13c)$$

whereas a *flux-controlled meminductor* is characterized by

$$\rho' = \varphi \quad (14a)$$

$$\varphi' = v \quad (14b)$$

$$i = L^{-1}(\rho)\varphi. \quad (14c)$$

Noteworthy, the relations (13c) and (14c) arise as the differentiated form of certain mappings  $\varphi = \gamma(\sigma)$  and  $q = \delta(\rho)$ , via the relations  $\sigma' = q, \rho' = \varphi$ . In a natural way this leads to other second order devices, such as those relating  $\sigma$  and  $\rho$  (cf. [5]). And in a fully nonlinear formalism, we may replace (13c) and (14c) by characteristics of the form

$$v = \nu(\sigma, q), \quad (15)$$

and

$$i = \chi(\rho, \varphi), \quad (16)$$

respectively.

In greater generality, one may look at the mathematical formalism of nonlinear circuit theory as a framework in which the basic circuit variables  $v$  (voltage) and  $i$  (current), together with some integral variables (not only  $\sigma$  and  $q$  but eventually higher order ones as well), are interrelated to define a dynamical system restricted only by Kirchhoff laws, which remain at (and actually define) the core of circuit theory. This general point of view is developed in [32], [33] and the reader is referred to these papers for details. See also [34] and the above-referenced paper [5].

## V. SOME APPLICATIONS

Many applications of memristive devices are reported in the design of nonlinear oscillators and chaotic circuits; we refer readers interested on this application area to the papers [4], [6], [9], [11], [12], [20], [21], [22], [27], [29] and references therein. Memristors are very promising in non-volatile memory design and in the implementation of resistive random access memory (RRAM or ReRAM); some recent references in this direction are [30], [35], [36], [37], [38], [39]. Digital logic applications are reported in [25], [28], [52], [53], [54] and references therein. There is already a vast amount of literature on memristive neural networks; cf. [16], [23], [25], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51]; this research activity stems from the fact that the memristor provides a very natural way to implement a synapse, displaying two stable states which are easily adjustable using electrical pulses. This approach paves the way for an efficient hardware implementation of artificial neural networks. Other actual or potential applications of memristors are discussed in [3], [5], [7], [8], [10], [13], [14], [15], [17], [18], [19], [24], [28], [52].

## VI. CONCLUDING REMARKS

Research on memristors and memristive devices has exploded in the last seven years, following the announcement by HP of the design of a nanoscale memristor in 2008. Traditional points of view on the fundamentals of circuit theory are being revisited and a lot of applications are being reported. Many analytical properties of memristive circuits remain to be solved, though, and much more applications are yet to come. This communication reports some recent research directions in this area, and the reader is referred to the references compiled below for a more detailed introduction to the main features and applications of memristive devices and systems.

## ACKNOWLEDGMENT

Research supported by Project MTM2010-15102 of Ministerio de Ciencia e Innovación, Spain.

## REFERENCES

- [1] L. O. Chua, Memristor – The missing circuit element, *IEEE Trans. Circuit Theory* **18** (1971) 507-519.
- [2] D. B. Strukov, G. S. Snider, D. R. Stewart and R. S. Williams, The missing memristor found, *Nature* **453** (2008) 80-83.
- [3] A. Ascoli, T. Schmidt, R. Tetzlaff and F. Corinto, Application of the Volterra series paradigm to memristive systems, in R. Tetzlaff (ed.), *Memristors and Memristive Systems*, pp. 163-191, Springer, 2014.
- [4] B. Bao, Z. Ma, J. Xu, Z. Liu and Q. Xu, A simple memristor chaotic circuit with complex dynamics, *Internat. J. Bifurcation and Chaos* **21** (2011) 2629-2645.
- [5] D. Birolek, Z. Birolek and V. Biolkova, SPICE modeling of memristive, memcapacitive and meminductive systems, *Proc. Eur. Conf. Circuit Theor. Design 2009*, pp. 249-252, 2009.
- [6] A. Buscarino, L. Fortuna, M. Frasca and L. V. Gambuzza, A chaotic circuit based on HewlettPackard memristor, *Chaos* **22** (2012) 023136.
- [7] F. Corinto, A. Ascoli and M. Gilli, Analysis of current-voltage characteristics for memristive elements in pattern recognition systems, *Internat. J. Circuit Theory Appl.* **40** (2012) 1277-1320.
- [8] M. Di Ventra, Y. V. Pershin and L. O. Chua, Circuit elements with memory: memristors, memcapacitors and meminductors, *Proc. IEEE* **97** (2009) 1717-1724.
- [9] M. Itoh and L. O. Chua, Memristor oscillators, *Internat. J. Bifur. Chaos* **18** (2008) 3183-3206.
- [10] M. Itoh and L. O. Chua, Memristor cellular automata and memristor discrete-time cellular neural networks, *Intl. J. Bifurcation and Chaos* **19** (2009) 3605-3656.
- [11] M. Itoh and L. O. Chua, Memristor Hamiltonian circuits, *Internat. J. Bifur. Chaos* **21** (2011) 2395-2425.
- [12] M. Itoh and L. O. Chua, Dynamics of memristor circuits, *Internat. J. Bifur. Chaos* **24** (2014) 1430015.
- [13] L. Jansen, M. Matthes and C. Tischendorf, Global unique solvability for memristive circuit DAEs of index 1, *Int. J. Circuit Theory Appl.*, in press, 2015.
- [14] D. Jeltsema and A. J. van der Schaft, Memristive port-Hamiltonian systems, *Math. Comp. Model. Dyn. Sys.* **16** (2010) 75-93.
- [15] D. Jeltsema and A. Doria-Cerezo, Port-Hamiltonian formulation of systems with memory, *Proc. IEEE* **100** (2012) 1928-1937.
- [16] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder and W. Lu, Nanoscale memristor device as synapse in neuromorphic systems, *Nano Letters* **10** (2010) 1297-1301.
- [17] O. Kavehei, A. Iqbal, Y. S. Kim, K. Eshraghian, S. F. Al-Sarawi and D. Abbott, The fourth element: characteristics, modelling and electromagnetic theory of the memristor, *Proc. R. Soc. A* **466** (2010) 2175-2202.
- [18] M. Messias, C. Nespole and V. A. Botta, Hopf bifurcation from lines of equilibria without parameters in memristors oscillators, *Internat. J. Bifur. Chaos* **20** (2010) 437-450.
- [19] B. Mouttet, A memadmittance systems model for thin film memory materials, preprint, arXiv:1003.2842, 2010.
- [20] B. Muthuswamy, Implementing memristor based chaotic circuits, *Internat. J. Bifur. Chaos* **20** (2010) 1335-1350.
- [21] B. Muthuswamy and L. O. Chua, Simplest chaotic circuit, *Internat. J. Bifur. Chaos* **20** (2010) 1567-1580.
- [22] B. Muthuswamy and P. P. Kokate, Memristor-based chaotic circuits, *IETE Tech. Rev.* **26** (2009) 417-429.
- [23] Y. V. Pershin and M. Di Ventra, Experimental demonstration of associative memory with memristive neural networks, *Neural Networks* **23** (2010) 881-886.
- [24] Y. V. Pershin and M. Di Ventra, Memory effects in complex materials and nanoscale systems, *Advances in Physics* **60** (2011) 145-227.
- [25] Y. V. Pershin and M. Di Ventra, Neuromorphic, digital and quantum computation with memory circuit elements, *Proc. IEEE* **100** (2012) 2071-2080.
- [26] R. Riaza, Nondegeneracy conditions for active memristive circuits, *IEEE Trans. Circuits and Systems - II* **57** (2010) 223-227.
- [27] W. Shen, Z. Zenga and F. Zou, A fractional-order chaotic circuit based on memristor and its generalized projective synchronization, in *Intelligent Computing Theory*, Lecture Notes in Computer Science 8588, Springer, 2014, pp 838-844.
- [28] R. Tetzlaff (ed.), *Memristors and Memristive Systems*, Springer, 2014.
- [29] S. Wen, Z. Zeng, T. Huang and Y. Chen, Fuzzy modeling and synchronization of different memristor-based chaotic circuits, *Physics Letters A* **377** (2013) 2016-2021.
- [30] J. J. Yang, M. D. Pickett, X. Li, D. A. Ohlberg, D. R. Stewart and R. S. Williams, Memristive switching mechanism for metal/oxide/metal nanodevices, *Nature Nanotechnology* **3** (2008) 429-433.
- [31] L. O. Chua and S. M. Kang, Memristive devices and systems, *Proc. IEEE* **64** (1976) 209-223.
- [32] R. Riaza, First order mem-circuits: modeling, nonlinear oscillations and bifurcations, *IEEE Trans. Circ. Sys. I* **60** (2013) 1570-1583.

- [33] R. Riaza, Second order mem-circuits, *Int. J. Circuit Theory Appl.*, in press, 2015.
- [34] L. O. Chua, Nonlinear circuit foundations for nanodevices, Part I: The four-element torus, *Proc. IEEE* **91** (2003) 1830-1859.
- [35] L. O. Chua, Resistance switching memories are memristors, *Applied Physics A* **102** (2011) 765-783.
- [36] F. T. Chen, Y. Chen, T. Y. Wu and T. K. Ku, Write scheme allowing reduced LRS nonlinearity requirement in a 3D-RRAM array with selector-less 1T1R architecture, *Electron Device Letters* **35** (2014) 223-225.
- [37] E. Gale, TiO<sub>2</sub>-based memristors and ReRAM: materials, mechanisms and models (a review), *Semicond. Sci. Technol.* **29** (2014) 104004.
- [38] A. Mehonic, S. Cueff, M. Wojdak, S. Hudziak, O. Jambois, C. Labbe, B. Garrido, R. Rizk and A.J. Kenyon, Resistive switching in silicon sub-oxide films, *J. Appl. Phys.* **111** 074507 (2012).
- [39] H. P. Wong, H. Y. Lee, S. Yu, Y. S. Chen, Y. Wu, P. Chen, B. Lee, F.T. Chen and M. Tsai, Metal-oxide RRAM, *Proc. IEEE* **100** (2012) 1951-1970.
- [40] S. P. Adhikari, C. Yang, H. Kim and L. O. Chua, Memristor bridge synapse-based neural network and its learning, *IEEE Trans. Neural Networks and Learning Systems* **23** (2012) 1426-1435.
- [41] I. E. Ebong and P. Mazumder, CMOS and memristor-based neural network design for position detection, *Proc. IEEE* **100** (2012) 2050-2060.
- [42] H. Kim, M. P. Sah, C. Yang, T. Roska and L. O. Chua, Neural synaptic weighting with a pulse-based memristor circuit, *IEEE Transactions on Circuits and Systems I: Regular Papers* **59** (2012) 148-158.
- [43] H. Kim, M. P. Sah, C. Yang, T. Roska and L. O. Chua, Memristor bridge synapse, *Proc. IEEE* **100** (2012) 2061-2070.
- [44] D. Querlioz, O. Bichler and C. Gamrat, Simulation of a memristor-based spiking neural network immune to device variations, *Proc. 2011 Intl. Joint Conf. on Neural Networks (IJCNN)* (2011), 1775 - 1781
- [45] M. P. Sah, C. Yang, H. Kim and L. O. Chua, A voltage mode memristor bridge synaptic circuit with memristor emulators, *Sensors* **12** (2012) 3587-3604.
- [46] T. Serrano-Gotarredona, T. Masquelier, T. Prodromakis, G. Indiveri and B. Linares-Barranco, STDP and STDP variations with memristors for spiking neuromorphic learning systems, *Frontiers in Neuroscience* **7** (2013) 2013-2.
- [47] D. Soudry, D. Di Castro, A. Gal, A. Kolodny and S. Kvatinisky, Memristor-based multilayer neural networks with online gradient descent training, CCIT Technical Report #840, 2013.
- [48] J. Z. Starzyk and Basawaraj, Comparison of two memristor based neural network learning schemes for crossbar architecture, *Proc. IWANN'2013*, Part I, 492-499 (2013).
- [49] A. Thomas, Memristor-based neural networks, *J. Phys. D: Appl. Phys.* **46** (2013) 093001.
- [50] S. Wen, Z. Zeng and T. Huang, Exponential stability analysis of memristor-based recurrent neural networks with time-varying delays, *Neurocomputing* **97** (2012) 233-240.
- [51] X. Wu, V. Saxena and K. A. Campbell, Energy-efficient STDP-based learning circuits with memristor synapses, *Proc. SPIE 9119, Machine Intelligence and Bio-inspired Computation: Theory and Applications VIII*, 9119-06 (2014).
- [52] A. Adamatzky and L. O. Chua (eds.), *Memristor Networks*, Springer, 2014.
- [53] H. Owlia, P. Keshavarzi and A. Rezaei, A novel digital logic implementation approach on nanocrossbar arrays using memristor-based multiplexers, *Microelectronics Journal* **45** (2014) 597-603.
- [54] T. Raja and S. Mourad, Digital logic implementation in memristor-based crossbars, *Proc. Intl. Conf. on Communications, Circuits and Systems 2009 (ICCCAS 2009)*, pp. 939-943, 2009.

**Ricardo Riaza** (Madrid, Spain, 1972) received the MSc degree in Mathematics from Universidad Complutense de Madrid in 1996, the MSc degree in Electrical and Electronic (Telecommunication) Engineering from Universidad Politécnica de Madrid (UPM) in 1997, and the PhD degree in Mathematics from UPM in 2000. He received the Outstanding PhD Award for his Doctoral Thesis in 2001 and the Research/Technological Development Award for young professors in 2005, both from UPM. Currently, he serves as an Associate Professor at the Departamento de Matemática Aplicada a las Tecnologías de la Información y las Comunicaciones of the ETSI Telecomunicación (UPM). His current research interests are focused on differential-algebraic equations (DAEs) and also on analytical aspects of nonlinear electrical and electronic circuits, including circuits with memristors. He is the author of the book "Differential-Algebraic Systems: Analytical Aspects and Circuit Applications" (World Scientific, 2008), and the first or unique author of nearly forty papers in JCR journals.

# Detection Singular Polarization State by Multi-Order Diffractive Optical Element

Dmitry A. Savelyev, Nikolay L. Kazanskiy, and Svetlana N. Khonina

**Abstract**—We investigate detection of polarization states of the focused incident beam using vortex phase elements. As a focusing system we use the multi-order diffractive optical element. It is numerically shown possibility to recognize singular polarization types (circular, radial and azimuthal).

**Keywords**—singular polarization state, optical vortices, multi-order diffractive optical element.

## I. INTRODUCTION

IN this article, we use vortex phase elements in order to research a possibility of detecting the focused incident beam polarization state. Their complex-valued function can be written as a superposition of optical vortices. The effect of mutual influence between optical phase vortices and polarization singularities is well studied. Both their transformation into each other and the angular momentum compensation or enhancement has been researched for a long period of time [1-13]. Therefore, the idea of using the vortex phase to analyze laser fields polarizing properties is rather obvious [14-16]. However, generally it is impossible to reveal the interrelation between scalar (phase) and vector (polarization) optical vortices visually, except in the case of a high numerical aperture (NA) mode (e.g., sharp focusing) being used [15-19].

The problem can be solved by the instrumentality of diffractive optics [20] superposed to the focusing system [15-19]. More than that, we can insert a singularity into a focusing element structure [21].

Using a micro-objective [22], a parabolic mirror [23, 24], a diffractive lens [24-26], or an axicon [27-31], we can get the sharp focusing. It was supposed [24] that a sharper (than in

the case of a micro-objective) focusing could be achieved by means of a parabolic mirror or a diffractive lens. The experiment showed it to be true for the parabolic mirror [23]. As for a diffractive lens, the proposition was confirmed numerically [25, 26]. In addition, the insertion of axicon structures was proved to enhance the aplanatic lens focusing properties [26, 32].

We analyze case of sharp focusing with high-numerical-aperture micro-objective supplemented by the multi-order diffractive optical element. Here we present the simulation results in the frame of the Debye approximation [33] and the plane wave expansion method [34]. Our results are very similar with results obtained by means of the diffractive axicon [35].

## II. MULTI-ORDER DIFFRACTIVE OPTICAL ELEMENT

From above researches clearly, that for unequivocal detection of polarization type there can be not enough one test vortex phase even at sharp focusing [36]. It is desirable to know simultaneously results of action several optical vortices and even their superposition. Besides the certain combinations of optical vortices are convenient from the practical point of view since it may be generated by means of simple binary phase elements [19, 37].

To realize the simultaneous response of an analyzed field for several vortices, it is possible to use multi-order binary optical elements [38-40]. The binary phase transmission function is shown in Fig. 1, which allowing to receive the response of an analyzed beam to various combinations of phase vortices simultaneously in several diffractive orders in

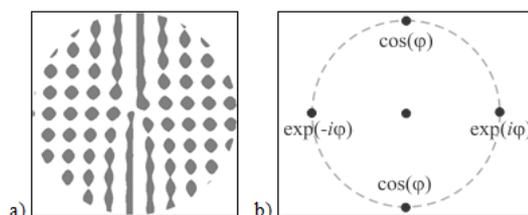


Fig. 1 transmission function of the focusing system: (a) binary phase of multi-order diffractive optical element, (b) accordance of diffractive orders to combinations of optical vortices

the focal plane.

Results of recognition of orthogonal linear polarization states are shown in Table 1 for cases of absence and presence

This work was supported by the Russian Foundation for Basic Research (grants 13-07-00266, 14-07-31079 mol\_a) and by the Ministry of Education and Science of Russian Federation.

D. A. Savelyev is with the Image Processing Systems Institute of the RAS, Samara, 443001, Russia; and the Samara State Aerospace University named after academician S.P. Korolyov, Samara, 443086, Russia (e-mail: dmitrey.savelyev@yandex.ru).

N. L. Kazanskiy is with the Image Processing Systems Institute of the RAS, Samara, 443001, Russia; and the Samara State Aerospace University named after academician S.P. Korolyov, Samara, 443086, Russia (phone: (846)332-57-83; e-mail: kazansky@smr.ru).

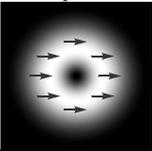
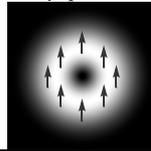
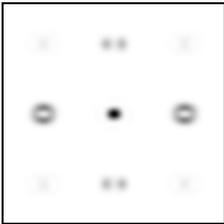
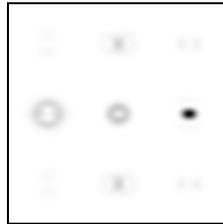
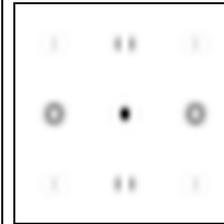
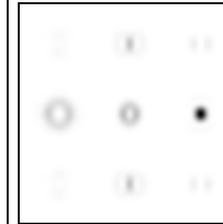
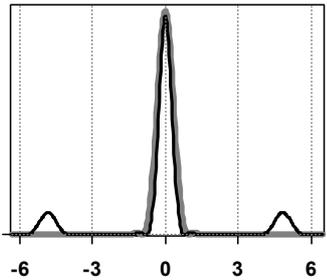
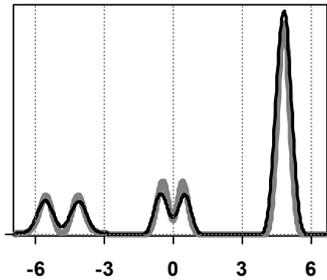
S. N. Khonina is with the Image Processing Systems Institute of the RAS, Samara, 443001, Russia; and the Samara State Aerospace University named after academician S.P. Korolyov, Samara, 443086, Russia (e-mail: khonina@smr.ru).

of vortex phase in an analyzed beam. The amplitude of the beam is a Gaussian function multiplied by radius. The phase of the beam is constant or vortex of the first order.

Parameters of calculation: length of a wave of incident radiation is  $\lambda = 1$  micrometer, a focal length  $f = 101$

micrometers, the numerical aperture of micro-objective  $NA = 0.99$ , radius of Gaussian beam is 50 micrometers; considered focal area is 15 micrometers  $\times$  15 micrometers.

Table I. Detection of the orthogonal linear polarization states for cases of absence and presence of vortex phase in an analyzed beam

|   |  |   |   |   |
|---|--|---|---|---|
| Analyzed beam   | Linear x-polarization  |   | Linear y-polarization   |   |
|   |   |   |  |   |
| Intensity distribution in focal plane (negative)  | with vortex  |   | with vortex   |   |
|   |   |  |   |  |
| Intensity section in focal plane (black line for x-polarization and grey line for y-polarization) | Vertical section without vortex:   |   | Horizontal section with vortex:   |   |
|   |  |   |  |   |

As seen from the modeling results presence or absence vortex phase of the first order is easily found out by presence of correlation peak in the corresponding diffractive order in the focal plane: absence of optical vortex corresponds to bright light intensity in the center of the focal plane.

Recognition of the orthogonal linear polarizations is carried out by less obvious characteristics. In particular, if the analyzed beam has no vortex phase then in vertical diffractive orders (corresponding to  $\cos(\varphi)$ ) will be nonzero value of intensity at x-polarization whereas for y-polarization in these points intensity will be absent (the comparative graph is in the third column of Table 1). At presence of the vortex phase in the analyzed beam recognition of polarization becomes very doubt.

Results of recognition of the orthogonal circular polarizations are shown in Table 2 in absence and at presence of vortex phase in an analyzed beam. Detection of vortex phase is similar to the previous case.

Recognition of the orthogonal circular polarizations is much easier than linear polarizations because in former case polarizing singularity is closely connected with phase singularity. We shall show it by presenting circular

polarization in polar components:

$$\mathbf{e}_x \pm i\mathbf{e}_y = (\mathbf{e}_r \cos \varphi - \mathbf{e}_\varphi \sin \varphi) \pm \pm i(\mathbf{e}_r \sin \varphi + \mathbf{e}_\varphi \cos \varphi) = \exp(\pm i\varphi)[\mathbf{e}_r \pm i\mathbf{e}_\varphi]. \quad (1)$$

As follows from (1) the circular polarization is corresponding to the vortex phase of the first order with the same sign as the direction of polarization.

The interrelation of the polarizing singularity with the vortex phase is obvious in nonzero intensity in corresponding vortex diffractive orders (Table 2) though there is no phase vortex in the analyzed beams.

When vortex phase is present in an analyzed beam it is also easy to distinguish type of polarization: if the directions of circular polarization and phase vortex are identical in the central diffractive order will be zero intensity. If the directions are opposite then in the center of the focal plane will be a nonzero value (the comparative graph is in the third column of Table 2).

Results of recognition of the orthogonal cylindrical polarizations are shown in Table 3 in absence and at presence of vortex phase in an analyzed beam.

Table II. Detection of the orthogonal circular polarization states for cases of absence and presence of vortex phase in an analyzed beam

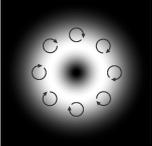
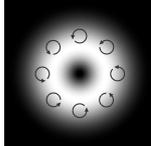
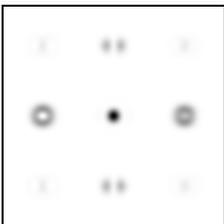
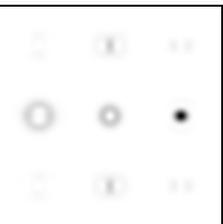
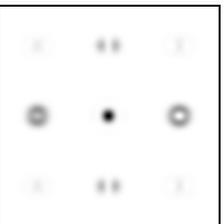
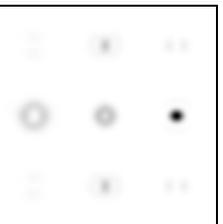
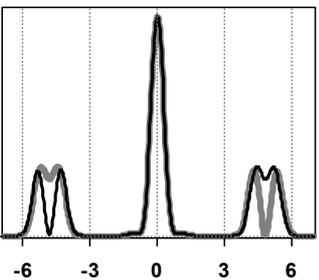
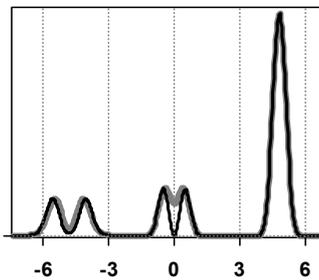
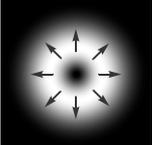
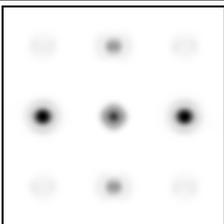
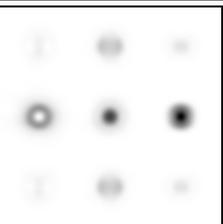
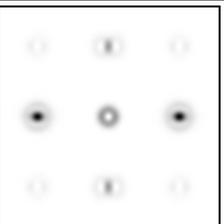
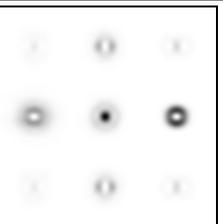
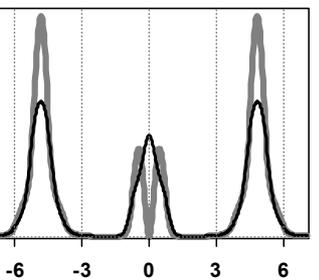
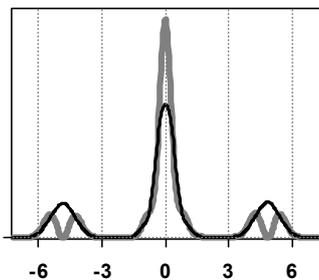
|   |  |   |   |   |
|---|--|---|---|---|
| Analyzed beam   | Circular «+»-polarization  |   | Circular «-»-polarization   |   |
|   |   |  |   |   |
| Intensity distribution in focal plane (negative)  | <b>with vortex</b>   |   | <b>with vortex</b>  |   |
|   |   |    |   |  |
| Intensity section in focal plane (black line for x-polarization and grey line for y-polarization) | Horizontal section without vortex:   |   | Horizontal section with vortex:   |   |
|   |  |   |  |   |

Table III. Detection of the orthogonal cylindrical polarization states for cases of absence and presence of vortex phase in an analyzed beam

|   |   |   |  |   |
|---|---|---|--|---|
| Analyzed beam   | Radial polarization   |   | Azimuthal polarization   |   |
|   |  |  |  |   |
| Intensity distribution in focal plane (negative)  | <b>with vortex</b>  |   | <b>with vortex</b>   |   |
|   |  |    |  |  |
| Intensity section in focal plane (black line for x-polarization and grey line for y-polarization) | Horizontal section without vortex:  |   | Vertical section with vortex:  |   |
|   |  |   |  |   |

Recognition of the orthogonal cylindrical polarizations is visually obvious because this type of polarization also is

singular and is connected with vortex phase. In particular, for radial and azimuthal polarization, accordingly:

$$\begin{aligned}
 \mathbf{e}_r &= \mathbf{e}_x \cos \varphi + \mathbf{e}_y \sin \varphi = \\
 &= \frac{1}{2} \exp(i\varphi) [\mathbf{e}_x - i\mathbf{e}_y] + \frac{1}{2} \exp(-i\varphi) [\mathbf{e}_x + i\mathbf{e}_y], \\
 (2) \quad \mathbf{e}_\varphi &= -\mathbf{e}_x \sin \varphi + \mathbf{e}_y \cos \varphi = \\
 &= \frac{1}{2} \exp(i\varphi) [\mathbf{e}_y + i\mathbf{e}_x] + \frac{1}{2} \exp(-i\varphi) [\mathbf{e}_y - i\mathbf{e}_x]. \quad (3)
 \end{aligned}$$

From expressions (2) and (3) it is following that a cylindrical polarization contains vortex phases of the first order of both signs. Such interrelation of the cylindrical polarization with vortex phase is well appreciable in nonzero intensity in corresponding vortex diffractive orders (Table 3) though there is no vortex phase in an analyzed beam.

In this case detection of vortex phase is defined in correlation peak in the central diffractive order with simultaneous absence of intensity in any vortex (horizontal) diffractive order.

At presence of vortex phase in an analyzed beam it is also easy to distinguish type of polarization: for radial polarization vertical diffractive orders have nonzero value, and for azimuthal they have zero intensity (the comparative graph is in the third column of Table 3).

### III. CONCLUSION

Carried out researches have shown that in conditions of sharp focusing the multi-order diffractive optical element matched with various combinations of vortex phase allows to distinguish unequivocally singular polarization states (circular, radial and azimuthal). Unambiguity of recognition is provided by interrelation of polarizing and phase singularities.

For linear polarization such interrelation is absent therefore detailed recognition of various types of linear polarizations by means of micro-objective focusing is complicated. In work [36] it has been shown that focusing by means of diffractive axicon allows to detect a polarization state with better results. It can be explain by the following: the NA of a micro-objective has various values depending on radius and reaches the maximal value only on the edge of an optical element. Therefore the beams going under various angles are crossed in the focus, and the information connected with the high NA is imposed by the information carrying by paraxial beams. An axicon has the identical NA at any radius - both in the center, and at edge, therefore the information corresponding to one value of the numerical aperture takes place in the focus.

### REFERENCES

- [1] R.A. Beth, "Mechanical detection and measurement of the angular momentum of light," *Phys. Rev.* vol. 50, pp. 115–125, 1936.
- [2] A.H.S. Holbourn, "Angular momentum of circularly polarized light," *Nature (London)* vol. 137, p. 31, 1936.
- [3] J.F. Nye and M.V. Berry, "Dislocations in wave trains," *Proc. R. Soc. London Ser. A*, vol. 336, p. 165, 1974.
- [4] M.V. Berry, "The adiabatic phase and Pancharatnam's phase for polarized light," *J. Mod. Opt.* vol. 34, pp. 1401–1407, 1987.
- [5] L. Allen, M.W. Beijersbergen, R.J.C. Spreeuw, and J.P. Woerdman, "Orbital angular momentum of light and the transformation of Laguerre-Gaussian laser modes," *Phys. Rev. A*, vol. 45, pp. 8185–8189 1992.
- [6] S.M. Barnett and L. Allen, "Orbital angular momentum and nonparaxial light beams," *Opt. Commun.*, vol. 110, pp. 670–678, 1994.
- [7] N.B. Simpson, K. Dholakia, L. Allen, and M.J. Padgett, "Mechanical equivalence of spin and orbital angular momentum of light: an optical spanner," *Optics Letters*, vol. 22(1), pp. 52–54, 1997.
- [8] L. Allen, M.J. Padgett, and M. Babiker, "The orbital angular momentum of light," *Progress Optics*, vol. 39, p. 291, 1999.
- [9] M. Soskin and M.V. Vasnetsov, "Singular optics," *Progress Optics*, vol. 42, p. 219, 2001.
- [10] A.S. Desyatnikov, L. Torner, and Y.S. Kivshar, "Optical vortices and vortex solitons," *Progress Optics*, vol. 47, pp. 219–319, 2005.
- [11] G. Molina-Terriza, J.P. Torres, and L. Torner, "Twisted photons," *Nature Phys.*, vol. 3, pp. 305–310, 2007.
- [12] S. Franke-Arnold, L. Allen, and M. Padgett, "Advances in optical angular momentum," *Laser Photonics Rev.*, vol. 2, pp. 299–313, 2008.
- [13] M. R. Dennis, K. O'Holleran, and M. J. Padgett, "Singular optics: optical vortices and polarization singularities," *Progress in Optics*, vol. 53, pp. 293–363, 2009.
- [14] J. Leach, J. Courtial, K. Skeldon, S.M. Barnett, S. Franke-Arnold, and M.J. Padgett, "Interferometric methods to measure orbital and spin, or the total angular momentum of a single photon," *Physical Review Letters*, vol. 92, no. 1, pp. 013601, 2004.
- [15] L.E. Helseth, "Optical vortices in focal regions," *Opt. Commun.*, vol. 229, pp. 85–91, 2004.
- [16] Y. Zhao, J.S. Edgar, G.D.M. Jeffries, D. McGloin, and D.T. Chiu, "Spin-to-orbital angular momentum conversion in a strongly focused optical beam," *Physical Review Letters*, vol. 99, pp. 073901, 2007.
- [17] I. Moreno, J.A. Davis, I. Ruiz, and D.M. Cottrell, "Decomposition of radially and azimuthally polarized beams using a circular-polarization and vortex-sensing diffraction grating," *Optics Express*, vol. 18, no. 7, pp. 7173–7183, 2010.
- [18] S.N. Khonina, N.L. Kazanskiy, and S.G. Volotovskiy, "Vortex phase transmission function as a factor to reduce the focal spot of high-aperture focusing system," *Journal of Modern Optics*, vol. 58, no. 9, pp. 748–760, 2011.
- [19] S.N. Khonina, "Simple phase optical elements for narrowing of a focal spot in high-numerical-aperture conditions," *Optical Engineering*, vol. 52, no. 9, pp. 091711, 2013.
- [20] V.A. Soifer, *Methods for Computer Design of Diffractive Optical Elements*, John Wiley & Sons Inc., 2002.
- [21] S.N. Khonina, D.V. Nesterenko, A.A. Morozov, R.V. Skidanov, and V.A. Soifer, "Narrowing of a light spot at diffraction of linearly-polarized beam on binary asymmetric axicons," *Optical Memory and Neural Networks (Information Optics)*, vol. 21, no. 1, pp. 17–26, 2012.
- [22] R. Dorn, S. Quabis, and G. Leuchs, "Sharper focus for a radially polarized light beam," *Physical Review Letters*, vol. 91, no. 23, pp. 233901, 2003.
- [23] J. Stadler, C. Stanciu, C. Stupperich, and A. J. Meixner, "Tighter focusing with a parabolic mirror," *Optics Letters*, vol. 33, no. 7, pp. 681–683, 2008.
- [24] N. Davidson and N. Bokor, "High-numerical-aperture focusing of radially polarized doughnut beams with a parabolic mirror and a flat diffractive lens," *Optics Letters*, vol. 29, no. 12, pp. 1318–1320, 2004.
- [25] N. Sergienko, V. Dhayalan, and J. J. Stamnes, "Comparison of focusing properties of conventional and diffractive lens," *Opt. Commun.*, vol. 194, pp. 225–234, 2001.
- [26] S.N. Khonina, and S.G. Volotovskiy, "Controlling the contribution of the electric field components to the focus of a high-aperture lens using binary phase structures," *J. Opt. Soc. Am. A*, vol. 27, no 10, pp. 2188–2197, 2010.
- [27] V. P. Kalosha and I. Golub, "Toward the subdiffraction focusing limit of optical superresolution," *Opt. Lett.*, vol. 32, pp. 3540–3542, 2007.
- [28] C.J.A. Zapata-Rodríguez Sánchez-Losa, "Three-dimensional field distribution in the focal region of low-Fresnel-number axicons," *J. Opt. Soc. Am. A*, vol. 23, no. 12, pp. 3016–3026, 2006.
- [29] V.V. Kotlyar, A.A. Kovalev, and S.S. Stafeev, "Sharp focus area of radially-polarized Gaussian beam propagation through an axicon," *Prog. Electromagn. Res. C*, vol. 5, pp. 35–43, 2008.

- [30] S.N. Khonina, D.A. Savelyev, P.G. Serafimovich, and I.A. Pustovoy, "Diffraction at binary microaxicons in the near field," *J. Opt. Technol.*, vol. 79, no. 10, pp. 22–29, 2012.
- [31] S.N. Khonina, S.V. Karpeev, S.V. Alferov, D.A. Savelyev, J. Laukkanen, and J. Turunen, "Experimental demonstration of the generation of the longitudinal E-field component on the optical axis with high-numerical-aperture binary axicons illuminated by linearly and circularly polarized beams," *J. Opt.*, vol. 15, pp. 085704, 2013.
- [32] S.N. Khonina, N.L. Kazanskiy, A.V. Ustinov, and S.G. Volotovskiy, "The lensacon: nonparaxial effects," *J. Opt. Technol.*, vol. 78, no. 11, pp. 724–729, 2011.
- [33] B. Richards and E. Wolf, "Electromagnetic diffraction in optical systems II. Structure of the image field in an aplanatic system," *Proc. Royal Soc. A*, vol. 253, pp. 358–379, 1959.
- [34] M.B. Vinogradova, O.V. Rudenko, and A.P. Sukhorukov, *Theory of Waves*, Nauka, Moscow, 1979.
- [35] S.N. Khonina and D.A. Savelyev, "High-aperture binary axicons for the formation of the longitudinal electric field component on the optical axis for linear and circular polarizations of the illuminating beam," *Journal of Experimental and Theoretical Physics*, vol. 117, no. 4, pp. 623–630, 2013.
- [36] S.N. Khonina, D.A. Savelyev, N.L. Kazanskiy, and V.A. Soifer, "Singular phase elements as detectors for different polarizations," *Proc. SPIE*, vol. 9066, pp. 90660A, 2013.
- [37] Y. Unno, T. Ebihara, and M.D. Levenson, "Impact of Mask Errors and Lens Aberrations on the Image Formation by a Vortex Mask," *J. Microlith. Microfab. Microsys.*, vol. 4, no. 2, pp. 023006, 2005.
- [38] V.V. Kotlyar, S.N. Khonina, and V.A. Soifer, "Light field decomposition in angular harmonics by means of diffractive optics," *Journal of Modern Optics*, vol. 45, no. 7, pp. 1495–1506, 1998.
- [39] V.V. Kotlyar and S.N. Khonina, "Multi-order diffractive optical elements to process data," in *Perspectives in Engineering Optics*, Ed. by K. Singh, V.K. Rastogi, Anita Publications, Delhi, pp. 47–56, 2003.
- [40] S.N. Khonina, V.V. Kotlyar, V.A. Soifer, K. Jefimovs, J. Turunen, "Generation and selection of laser beams represented by a superposition of two angular harmonics," *Journal of Modern Optics*, vol. 51, no. 5, pp. 761–773, 2004.

# The structural constant of an atom as the basis of some known physical constants

Milan Perkovic

**Abstract**—Maxwell's equations are used in atomistic since it has been shown that they in addition to the emission of radiation include the simultaneous absorption and thereby maintain the stability of the atom. Conditions of existence of electromagnetic oscillation in the atom are met if the product of one part of the characteristic impedance of oscillator, called structural coefficient, and atomic number  $Z$  is constant. The structural constant of atom  $s_0 \approx 8.278$  was determined through the stability of atoms and through the ionization energy of the hydrogen atom. The fine-structure constant is calculated from this, namely  $s_0$  squared times two is inverse fine-structure constant 137.073. The meaning of the fine-structure constant is being discussed and proposes the use of these constants related to the isotope number. For this purpose, it is proposed to introduce a new unit. Planck's constant is obtained by multiplying the speed of light, magnetic constant, and charge of the electron and structural constant of atom squared. In a similar way, with the aid of structural constant of atom, Josephson's constant, von Klitzing's and Rydberg's constant were determined. These constants are here derived theoretically and are consistent with the measurements.

**Keywords**—Fine-structure constant, Josephson constant, Planck constant, Structural constant of atom, von Klitzing constant.

## I. INTRODUCTION

THE application of Maxwell's equations in the last hundred years has experienced a failure during the study phenomena in microcosm. Nevertheless, there are still tendencies of researchers to apply these equations except both, in the macro and micro world. The reasons for this tendency and for this paper as well include the fact that Maxwell's equations themselves do not possess a certain limit of their applicability. The main theoretical objection against the application of Maxwell's equations in the atomistic was that these equations require radiation of electrons, because there's acceleration of the charge on a circular orbit in an atom, causing the collapse of the atom. Since in reality the atoms do not collapse, the conclusion that Maxwell's equations are not valid in the atomistic seemed exactly right. One option has not been taken into account so far. It has been shown that the application of Maxwell's equations does not cause a collapse of atoms if these equations are observed completely, i.e., that

This work was supported in part by *Drives-Control*, Ltd., Zagreb, Croatia, [www.drivesc.com](http://www.drivesc.com) and *Prvomajska TZR*, Ltd., Zagreb, Croatia, [www.prvomajska-tzr.hr](http://www.prvomajska-tzr.hr).

Milan Perkovic is with the *First Technical School TESLA*, Klaićeva 7, Zagreb, Croatia (corresponding author to provide phone:+385-(0)-98-218-051; e-mail: [milan@drivesc.com](mailto:milan@drivesc.com)).

in addition to simultaneous emission and absorption of radiation there [1], [2]. Thus an atom with the application of Maxwell's equations remains stable. After that, there is no theoretical limit to the application of Maxwell's equations in the atom. Thus the atom can be treated as an electromechanical oscillator.

Using Maxwell's equations reveals the existence of a single constant of the atom that is made up of two structural parts, i.e., structural coefficient of Lecher's line  $\sigma(\chi)$  and the atomic number  $Z$ . One of these parts is bonded to the structure of Lecher's line, which serves as a model of an electromagnetic wave in an atom, and the other one is related to the atomic number, which is part of an atom. So we called it *the structural constant of an atom* (shorter, *structural constant*) and we marked it with  $s_0$ .

## II. A STRUCTURAL CONSTANT AND ITS CALCULATION

In the Section IV, I will show the origin of a structural atom constant. Now let's say that it is defined as [3]

$$s_0 \equiv \sqrt{\sigma(\chi) Z}, \quad (1)$$

where  $s_0$  is a structural atom constant,  $\sigma(\chi)$  is a structural coefficient of Lecher's line [Lecher line is twin-lead transmission line consisting of a pair of ideal conductive nonmagnetic wires of diameter  $2\rho$ , separated by  $\delta$ , situated in space with permittivity  $\epsilon$  and permeability  $\mu$ , whereby the argument  $\chi = \delta/\rho$ , and where the behavior of electromagnetic quantities as well as in the atom, in other words the Lecher line in our case represents the model of electromagnetic energy in the atom],  $Z$  is an atomic number, which theoretically is not in integer domain. The capacitance and the inductance of Lecher's line per unit length are

$$C' = \frac{\pi\epsilon}{\ln\left(\chi/2 + \sqrt{\chi^2/4 - 1}\right)} \quad (2)$$

and

$$L' = \frac{\mu(\ln\chi + 1/4)}{\pi}, \quad (3)$$

respectively. The characteristic impedance of Lecher's line is [4]

$$Z_{LC} = \sqrt{\frac{L}{C}} = \sqrt{\frac{L'\Delta z}{C'\Delta z}} = \sqrt{\frac{L'}{C'}} = \sqrt{\frac{\mu}{\varepsilon}} \frac{\sigma(\chi)}{\pi}. \quad (4)$$

It follows by using (1), [3],

$$\sigma(\chi) = \sqrt{\left[ \ln\left(\frac{\chi}{2} + \sqrt{\frac{\chi^2}{4} - 1}\right) \right] \left( \ln \chi + \frac{1}{4} \right)} = \frac{s_0^2}{Z}. \quad (5)$$

Both guides in Lecher's line are set parallel in  $z$ -axis. The wave of current ( $i$ ) and the wave of voltage ( $u$ ), i.e., the current-voltage waves of Lecher's line, taking place in time  $t$  along the  $z$ -axis. These two mutually connected waves are described with two differential equations, [4]:

$$\frac{\partial^2 u}{\partial z^2} - L'C' \frac{\partial^2 u}{\partial t^2} = 0, \quad \frac{\partial^2 i}{\partial z^2} - L'C' \frac{\partial^2 i}{\partial t^2} = 0. \quad (6)$$

On the other hand the electromagnetic wave (with  $x$ -component of electric field strength  $E_x$  and  $y$ -component of magnetic field strength  $H_y$ ) travels in general, but of course also within the atom, along the  $z$ -axis. This wave in the plane is also described with two differential equations, [4]:

$$\frac{\partial^2 E_x}{\partial z^2} - \varepsilon_0 \mu_0 \frac{\partial^2 E_x}{\partial t^2} = 0, \quad \frac{\partial^2 H_y}{\partial z^2} - \varepsilon_0 \mu_0 \frac{\partial^2 H_y}{\partial t^2} = 0. \quad (7)$$

From (6) and (7), we can see that the waves on the Lecher's line, and the electromagnetic waves in the atom, are described with the same form of differential equations. This means that all phenomena described by these equations are analogous. Therefore, it is possible to adapt the equations of current and voltage waves on Lecher's line, with the help of one factor  $F$ , so that it describes an electromagnetic wave in the atom, e.g., from (6) and (7) we obtain third equation, in this way,

$$\begin{aligned} \frac{\partial^2 u}{\partial z^2} - L'C' \frac{\partial^2 u}{\partial t^2} = 0, \quad \frac{\partial^2 E_x}{\partial z^2} - \varepsilon_0 \mu_0 \frac{\partial^2 E_x}{\partial t^2} = 0, \\ \frac{\partial^2 u}{\partial z^2} - F^2 L'C' \frac{\partial^2 u}{\partial t^2} = 0, \end{aligned} \quad (8)$$

which means that by multiplying  $L'C'$  by a factor  $F^2$  wave on Lecher's line behaves according to the phase velocity like electromagnetic wave in an atom. Therefore worth [as follows from (8)],

$$F^2 L'C' = \varepsilon_0 \mu_0, \quad (9)$$

or

$$F = \sqrt{\frac{\varepsilon_0 \mu_0}{\varepsilon \mu} \frac{\ln\left(\chi/2 + \sqrt{\chi^2/4 - 1}\right)}{\ln \chi + 1/4}}, \quad (10)$$

as gives in the case of Lecher's line within the vacuum ( $\varepsilon = \varepsilon_0, \mu = \mu_0$ ):

$$F(\chi) = \sqrt{\frac{\ln\left(\chi/2 + \sqrt{\chi^2/4 - 1}\right)}{\ln \chi + 1/4}}. \quad (11)$$

This further means that the phase velocity of the electromagnetic wave in an atom,  $u_{em}$ , can be expressed as

$$\begin{aligned} \frac{1}{u_{em}^2} &= [F(\chi)]^2 L'C' = \varepsilon_0 \mu_0 = \frac{1}{c^2}, \\ u_{em} &= \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = \lambda_{em} \nu_{em} = c \end{aligned} \quad (12)$$

while the phase velocity of the wave of current and voltage at the Lecher's line,  $u_{CV}$ , using (2), (3) and (11), can be expressed as (with  $\varepsilon = \varepsilon_0, \mu = \mu_0$ , which is an assumption we will continue to count with)

$$\frac{1}{u_{CV}^2} = L'C' = \frac{\varepsilon \mu (\ln \chi + 1/4)}{\ln\left(\chi/2 + \sqrt{\chi^2/4 - 1}\right)}, \quad (13)$$

$$\begin{aligned} u_{CV} &= \sqrt{\frac{\ln\left(\chi/2 + \sqrt{\chi^2/4 - 1}\right)}{\varepsilon_0 \mu_0 (\ln \chi + 1/4)}} \\ &= \lambda_{CV} \nu_{CV} = c F(\chi), \end{aligned} \quad (14)$$

where  $c = 1/\sqrt{\varepsilon_0 \mu_0}$  is the speed of light in vacuum,  $\lambda_{CV}$  is the wavelength of the wave of current and voltage on the Lecher's line, and the  $\nu_{CV}$  is the frequency of the same wave. Each electromagnetic wave in the atom thus is associated with one wave of current and voltage of Lecher's line. With that, following the behavior of waves on Lecher's lines we follow the behavior of the waves in the atom. Thus we connect the phase relationships of the waves on the Lecher's line and electromagnetic wave in the atom.

Equation (5) connects each specific atomic number  $Z$  to a specific argument  $\chi$ , for which the wave of current and voltage on Lecher's line is analogous with phase velocity of the electromagnetic wave in that atom [5]. In theoretical considerations  $Z$  is treated here as a continuous physical quantity which in reality takes discrete integer values. Later

we will connect electromagnetic energy of Lecher's line with the energy of the electromagnetic wave in the atom.

There may be at least three methods of determining the structural constant  $s_0$ .

- The first method is linked with the just treated phase velocity of the electromagnetic wave in the atom and the stability of atoms, where we need only one measured data (the atomic number  $Z$  of the first unstable atoms) [4]. In other words, this way answers the question, what is structural constant of atoms if the first unstable atom has atomic number  $Z$ . The percentage uncertainty of this method is 0.3%.

- The second method is connected to the ionization energy of hydrogen, where we need four data (ionization voltage  $V$ , electron charge  $e$ , electron mass  $m$  and the speed of light in vacuum  $c$ ). The relative standard uncertainty is  $1.3 \times 10^{-8}$ . It is three times better than that of Planck's constant, where the relative standard uncertainty, according to the NIST, is  $4.4 \times 10^{-8}$ .

- The third method is an extension of the second. Instead of the energies we use frequencies or wavelengths of spectral lines of atoms. Here we need six data (wavelength of the radiation of atoms  $\lambda$ , electron charge  $e$ , electron mass  $m$ , speed of light in vacuum  $c$ , magnetic constant  $\mu_0$  and the initial velocity of electrons  $\beta_0$ ). The relative standard uncertainty is therefore less favorable than in the second case, so we will not implement this method here.

### III. CALCULATION OF A STRUCTURAL CONSTANT THROUGH THE STABILITY OF ATOMS

The first method for the calculation of  $s_0$  is based on analysis of the stability of atoms [6]. It was concluded that the phase velocity  $u_{cv}$  on Lecher's line decreases towards zero when the argument  $\chi = \chi_0 = 2.328788$  or less, [4], Fig. 1. In this case the second derivative (in absolute value) becomes greater than 1, indicating a sharp drop the phase velocity, which leads to instability of the waves on the Lecher's line, that of the associated electromagnetic wave in the atom, and thus to instability of the whole atom.

The argument  $\chi_0$  therefore corresponds to the first unstable atom, and in 2003 it was found that it is bismuth, a chemical element with symbol Bi and atomic number 83, i.e.,  $^{83}\text{Bi}$ . The beginning of the instability series of the atoms may belong to the lower or upper part of the number 83, i.e.,  $Z=83 \pm 1/2$ . Using the above  $Z$  and  $\sigma(\chi_0) = \sigma(2.328788) = 0.825402$ , we get from (1) the result  $s_0 = 8.277 \pm 0.025$ , which means that percentage uncertainty is 0.3%. The validity of this result will be confirmed later by using NIST Atomic Spectra Database.

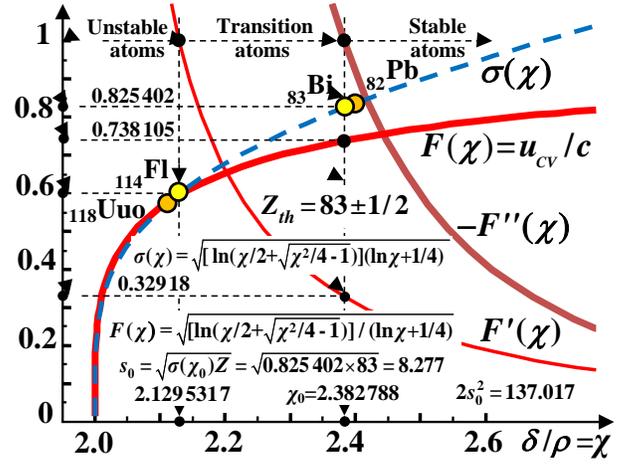


Fig. 1 Structural coefficient of Lecher's line  $\sigma(\chi)$ , normalized phase velocity of current and voltage of Lecher's line  $F(\chi)=u_{cv}/c$ , the first derivative of the normalized phase velocity  $F'(\chi)$ , inverted second derivative of the normalized phase velocity  $F''(\chi)$ , all versus ratio  $\delta/\rho=\chi$  of the transmission Lecher's line, consisting of a pair of ideal conducting nonmagnetic parallel wires of radius  $\rho$  separated by  $\delta$ .

### IV. CALCULATION OF A STRUCTURAL CONSTANT THROUGH THE IONIZATION ENERGY OF THE HYDROGEN

In similar manners as a phase velocities we also associate electromagnetic energy of Lecher's line and the energy of the electromagnetic wave in the atom. In this sense, we require that the energy of an electromagnetic wave in an atom is equal to the electromagnetic energy of Lecher's line.

The electromagnetic energy in the atom is obtained by using the energy balance. Newton's second law and Coulomb's law together give [7]

$$\frac{m v^2}{r \sqrt{1-\beta^2}} = \frac{|qQ|}{4\pi\epsilon_0 r^2}, \quad (15)$$

which gives

$$r = \frac{|qQ|}{4\pi\epsilon_0 m c^2} \frac{\sqrt{1-\beta^2}}{\beta^2}, \quad (16)$$

where  $r$  stands for the radius of the circular orbit of the electron,  $q$  is the charge of the electron ( $q=-e$ ),  $e$  is elementary charge,  $Q$  is the charge of the nucleus ( $Q=Ze$ ),  $m$  is the electron rest mass,  $v$  is the velocity of the electron,  $\beta = v/c$ ,  $\epsilon_0$  is permittivity of free space,  $\mu_0$  is permeability of free space,  $m/\sqrt{1-\beta^2}$  is the transverse mass of the electron [8]. Increasing transverse mass of the electron is  $\Delta m = m/\sqrt{1-\beta^2} - m$ . The kinetic energy of an electrons, [7],

$$K = \Delta m c^2 = \frac{m c^2}{\sqrt{1-\beta^2}} - m c^2, \quad (17)$$

and its potential energy [7], using (16),

$$U = \frac{qQ}{4\pi\epsilon_0 r} = -\frac{Ze^2}{4\pi\epsilon_0 r} = -\frac{mc^2}{\sqrt{1-\beta^2}} \beta^2. \quad (18)$$

A point charge  $Q$  created at a distance  $r$  from the charge (relative to the potential at infinity) the electric potential  $\Phi$ :

$$\Phi = \frac{Q}{4\pi\epsilon_0 r}, \quad (19)$$

so that the potential energy according to (18) can be written

$$U = q\Phi. \quad (20)$$

If we know the amount of potential energy  $U$  then we can determine  $\beta$  from (18):

$$\beta = \frac{1}{\sqrt{2} mc^2} \sqrt{\sqrt{U^4 + (2mc^2U)^2} - U^2}. \quad (21)$$

The total mechanical energy of an electron ( $W$ ) is the sum of its kinetic and potential energy [7]:

$$W = K + U = -mc^2 \left(1 - \sqrt{1 - \beta^2}\right). \quad (22)$$

If we know the total mechanical energy  $W = -eV$ , then from (21) we can also determine  $\beta$ :

$$\beta = \sqrt{1 - \left(1 - \frac{eV}{mc^2}\right)^2}, \quad (23)$$

where  $V$  is the potential difference through which the electron passes to get an equal energy as electromagnetic energy  $E_{em}$  or  $\Delta E_{em}$ . This energy corresponds to the energy of a photon.

According to the law of conservation of energy the energy of an isolated system, which includes the total energy of electrons  $W$  and electromagnetic energy of the atom  $E_{em}$ , is constant, despite internal changes (the energy disappearing in one form and reappearing in another or is transferred from one object to another):

$$W + E_{em} = const., \quad (24)$$

i.e.,

$$\Delta W + \Delta E_{em} = 0, \quad (25)$$

and using (22) we get

$$\begin{aligned} \Delta W &= \int_{\beta_0}^{\beta_n} dW = \int_{\beta_0}^{\beta_n} \left( -\frac{mc^2 \beta}{\sqrt{1-\beta^2}} \right) d\beta \\ &= -mc^2 \left( \sqrt{1-\beta_0^2} - \sqrt{1-\beta_n^2} \right) \\ &= -eV = -\Delta E_{em}, \end{aligned} \quad (26)$$

where  $\beta_0$  is the initial velocity and  $\beta$  (or  $\beta_n$ ) is a final velocity of the electron ( $\beta_0 \leq \beta_n$ ). From (26) we get

$$\sqrt{1-\beta_n^2} = \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2} \quad (27)$$

or

$$\begin{aligned} \beta_n^2 &= 1 - \left( \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2} \right)^2 \\ &= \frac{2\Delta E_{em}}{mc^2} \left( \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2 \beta_0^2}{2\Delta E_{em}} \right), \end{aligned} \quad (28)$$

which means it is

$$\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2 \beta_0^2}{2\Delta E_{em}} = \frac{mc^2 \beta_n^2}{2\Delta E_{em}}. \quad (29)$$

Using (27) and (28), (16) becomes

$$r = \frac{\frac{1}{2} \frac{|qQ|}{4\pi\epsilon_0 \Delta E_{em}} \left( \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2} \right)}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2 \beta_0^2}{2\Delta E_{em}}}, \quad (30)$$

or using (18), (29) and (30)

$$\Delta E_{em} = \frac{\frac{1}{2} \frac{|qQ|}{4\pi\epsilon_0 r} \left( \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2} \right)}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2 \beta_0^2}{2\Delta E_{em}}} \quad (31)$$

and from (29) and (31) we get

$$|U| = \frac{mc^2 \beta_n^2}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2}}. \quad (32)$$

In addition to equalizing the phase velocity of the waves in the atom and the waves on the Lecher's line, will also equalize the energy  $\Delta E_{em}$  of these waves. Let's find now an expression for wave energy on the Lecher's line. Lecher's line can be represented as an inductive-capacitive network (so-called  $LC$  network), which finally makes the oscillatory circuit ( $LC$  circuit) [9]. The natural frequency  $\nu_{LC}$  of  $LC$  circuit is [7]

$$\nu_{LC} = \frac{1}{2\pi\sqrt{LC}}, \quad (33)$$

where  $C$  is the sum of all small capacitances of the  $LC$  network on the open end of the network, and  $L$  is the sum of all small inductances of the  $LC$  network on the short-circuited of the network [9]. This frequency is equal to the frequency  $\nu$  of the electromagnetic wave generated in the atom

$$\nu = \nu_{LC}. \quad (34)$$

Electromagnetic energy in the  $LC$  circuit  $E_{LC}$  is equal to the maximum amount of energy on the inductor with an inductance  $L$ , or energy on the capacitor with a capacitance  $C$  [7],

$$E_{LC} = \frac{1}{2} \frac{\Theta^2}{C} = \Delta E_{em}, \quad (35)$$

where  $\Theta$  stands for maximal charge on the mentioned capacitor with the capacitance  $C$ . Using (31) and (35) we obtain

$$\frac{1}{2} \frac{\Theta^2}{C} = \frac{1}{2} \frac{|qQ|}{4\pi\epsilon_0 r} \left( \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2} \right) \cdot \quad (36)$$

$$\frac{1}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2\beta_0^2}{2\Delta E_{em}}}$$

One equation (36) has two unknowns,  $\Theta$  and  $C$ . By using Diophantine equations we get next couple of the many solutions [10]:

$$C = 4\pi\epsilon_0 r, \quad (37)$$

$$\Theta^2 = \frac{|qQ| \left[ \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2} \right]}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2\beta_0^2}{2\Delta E_{em}}}. \quad (38)$$

Equation (35), which represents the energy of the oscillatory  $LC$  circle, or the energy of the electromagnetic wave in the atom, can be written by using (33), (34) and (38):

$$\begin{aligned} \frac{1}{2} \frac{\Theta^2}{C} &= \frac{1}{2} \frac{\pi}{\pi} \frac{\Theta^2}{\sqrt{C}\sqrt{C}} \frac{\sqrt{L}}{\sqrt{L}} \\ &= \pi \sqrt{\frac{L}{C}} \Theta^2 \frac{1}{2\pi\sqrt{LC}} \\ &= \pi Z_{LC} \Theta^2 \nu = A \nu \end{aligned} \quad (39)$$

$$= \Delta E_{em} = \frac{\pi Z_{LC} |qQ| \left( \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2} \right)}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2\beta_0^2}{2\Delta E_{em}}} \nu,$$

where

$$A = \frac{\pi Z_{LC} |qQ| \sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2}}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2\beta_0^2}{2\Delta E_{em}}} \quad (40)$$

is the action of electromagnetic oscillator. One part of it which does not depend on  $\beta_0$  or on  $\Delta E_{em}$  can be denoted by  $A_0$ :

$$A_0 = \pi Z_{LC} |qQ|, \quad (41)$$

thus applies

$$A = A_0 \frac{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2}}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2\beta_0^2}{2\Delta E_{em}}}. \quad (42)$$

From (4) (i.e.,  $Z_{LC} = \sqrt{L/C}$ ), (32) and (33) [i.e.,  $\nu_{LC} = \nu = 1/(2\pi\sqrt{LC})$ ] we obtain  $\nu = 1/(2\pi Z_{LC} C)$ . Using (37) and the last part of (4) (provided that  $\epsilon = \epsilon_0$ ,  $\mu = \mu_0$ ,  $c = 1/\sqrt{\epsilon_0\mu_0}$ ) we get:

$$\nu = \frac{c}{8\pi\sigma(\chi) r}. \quad (43)$$

As we can see from (43) frequency  $\nu$  depends on the speed of light  $c$  and two spatial parameters,  $r$  and  $\sigma(\chi)$ . Note that this frequency is not in any way dependent on the charges. Inserting  $r$  from (30) into (43) gives:

$$\nu = \frac{\Delta E_{em}}{\mu_0 c \sigma(\chi) |qQ|} \frac{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{\beta_0^2 mc^2}{2\Delta E_{em}}}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2}}. \quad (44)$$

The equation (44) is dependent now on the charge  $|qQ|$ . Such a dependence of the charge, in accordance with (43), therefore

not exists. The physical quantity which has the dimension of the charge may be in expression (44), but its amount cannot be changed. Thus, (44) should be made independent of the charge. Such "neutralizing" of this equation can be achieved by converting the product  $\sigma(\chi)|qQ|$  in a constant, let's call it *charge constant of the structure*,  $q_0$ , in the following way:

$$\sigma(\chi)|qQ| = q_0^2. \quad (45)$$

Because

$$|qQ| = Ze^2, \quad (46)$$

applies

$$\sigma(\chi)Z = \left(\frac{q_0}{e}\right)^2 = s_0^2. \quad (47)$$

Thus we derived equation (1), i.e.,  $s_0 = \sqrt{\sigma(\chi)Z}$ , as promised at the beginning. From (4), (41) and (47) follows:

$$\begin{aligned} A_0 &= \pi Z_{LC} |qQ| = \sqrt{\frac{\mu_0}{\epsilon_0}} \sigma(\chi) Z e^2 \\ &= \sqrt{\frac{\mu_0}{\epsilon_0}} s_0^2 e^2 = c \mu_0 s_0^2 e^2. \end{aligned} \quad (48)$$

Equation (44), using (48), we can write:

$$\Delta E_{em} = A_0 v \frac{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2}}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2 \beta_0^2}{2\Delta E_{em}}}. \quad (49)$$

Combining (29) and (49) yields

$$\Delta E_{em} = mc^2 \left( \sqrt{1-\beta_0^2} - \frac{1}{2} \frac{mc^2 \beta_n^2}{A_0 v} \right). \quad (50)$$

Arranging and solving of (49) leads to

$$\begin{aligned} \Delta E_{em} &= eV \\ &= A_0 v + mc^2 \sqrt{1-\beta_0^2} - \sqrt{(A_0 v)^2 + (mc^2)^2}. \end{aligned} \quad (51)$$

From (35), (39) and (50) follows

$$\Delta E_{em} = Av, \quad (52)$$

and thence

$$A = A_0 + \frac{mc^2 \sqrt{1-\beta_0^2}}{v} - \sqrt{A_0^2 + \left(\frac{mc^2}{v}\right)^2}. \quad (53)$$

By organizing (44), using (45), (46) and (47), we obtain:

$$v = \frac{\Delta E_{em}}{A_0} \frac{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{2mc^2} + \frac{mc^2 \beta_0^2}{2\Delta E_{em}}}{\sqrt{1-\beta_0^2} - \frac{\Delta E_{em}}{mc^2}}, \quad (54)$$

which can be written in the form of extended Duane-Hunt's law,

$$v = \frac{eV}{A_0} \frac{\sqrt{1-\beta_0^2} - \frac{eV}{2mc^2} + \frac{mc^2 \beta_0^2}{2eV}}{\sqrt{1-\beta_0^2} - \frac{eV}{mc^2}}, \quad (55)$$

or with use of (40) and (41)

$$v = \frac{eV}{A}. \quad (56)$$

From (27), (28) and (55) follows

$$A_0 v = \frac{1}{2} \frac{mc^2 \beta^2}{\sqrt{1-\beta^2}} = \frac{1}{2} \frac{m}{\sqrt{1-\beta^2}} v^2, \quad (57)$$

which according to (18) gives

$$v = \frac{|U|}{2A_0}. \quad (58)$$

The solution of (57) gives  $\beta$  if we know  $A_0$ ,  $m$ ,  $c$  and  $v$ :

$$\beta = \frac{\sqrt{2}}{mc^2} \sqrt{\sqrt{(A_0 v)^4 + (mc^2 A_0 v)^2} - (A_0 v)^2}, \quad (59)$$

which together with (58) is the same as (21).

The momentum of the electromagnetic wave in an atom  $p_{em}$  is defined as the momentum of photon [7], [8], which means that it is equal to the ratio of energy  $\Delta E_{em}$  and the phase velocity  $u_{em}$  of electromagnetic wave in the atom, i.e.,

$$u_{em} = \lambda v, \quad (60)$$

whereby  $\lambda$  is wavelength of the electromagnetic wave in an atom. So momentum of electromagnetic wave in the atom reads

$$p_{em} = \frac{\Delta E_{em}}{u_{em}} = \frac{\Delta E_{em}}{\lambda \nu}, \quad (61)$$

and using (38) this expression becomes

$$p_{em} = \frac{A\nu}{\lambda \nu} = \frac{A}{\lambda}. \quad (62)$$

In accordance with the law of conservation of momentum, the momentum of the electromagnetic wave  $p_{em}$  is equal to the relativistic momentum of the electron [7],  $p = mv / \sqrt{1-v^2/c^2}$ ,

$$\frac{A}{\lambda} = \frac{mv}{\sqrt{1-\beta^2}} = \frac{mc\beta}{\sqrt{1-\beta^2}}. \quad (63)$$

With the use of (27), (28), (40) and (41) from (63) we get (with  $\Delta E_{em} = eV$ ):

$$\begin{aligned} \lambda &= \frac{A}{mv} \sqrt{1-\beta^2} = \frac{A}{mc} \frac{\sqrt{1-\beta^2}}{\beta} \\ &= \frac{A_0}{\sqrt{2meV}} \frac{\left( \sqrt{1-\beta_0^2} - \frac{eV}{mc^2} \right)^2}{\sqrt{\sqrt{1-\beta_0^2} - \frac{eV}{2mc^2} + \frac{mc^2\beta_0^2}{2eV}}}. \end{aligned} \quad (64)$$

The phase velocity of electromagnetic wave in an atom, according to (55), (60) and (64):

$$u_{em} = \sqrt{\frac{eV}{2m}} \frac{\sqrt{1-\beta_0^2} - \frac{eV}{mc^2}}{\sqrt{\sqrt{1-\beta_0^2} - \frac{eV}{2mc^2} + \frac{mc^2\beta_0^2}{2eV}}}. \quad (65)$$

From (42), (62) and (64) follows

$$p_{em} = \sqrt{2meV} \frac{\sqrt{\sqrt{1-\beta_0^2} - \frac{eV}{2mc^2} + \frac{mc^2\beta_0^2}{2eV}}}{\sqrt{1-\beta_0^2} - \frac{eV}{mc^2}}. \quad (66)$$

Using (30), (64) and (65) we obtain

$$\begin{aligned} \frac{\lambda}{r} &= \frac{8\pi\epsilon_0 A_0}{|qQ|} \sqrt{\frac{eV}{2m}} \frac{\sqrt{1-\beta_0^2} - \frac{eV}{mc^2}}{\sqrt{\sqrt{1-\beta_0^2} - \frac{eV}{2mc^2} + \frac{mc^2\beta_0^2}{2eV}}} \\ &= \frac{8\pi\epsilon_0 A_0}{|qQ|} u_{em}. \end{aligned} \quad (67)$$

From (18) and (67) follows

$$\lambda = 2A_0 \frac{u_{em}}{|U|}. \quad (68)$$

On the other hand, the same equation (68) we also obtain using (58) and (60).

At least two separate oscillatory processes are simultaneously performed within atoms. It is on the one hand uniform circular motion of electrons around the nucleus, as well as with other the oscillation electromagnetic energy generated within atoms, which acts as an electromagnetic wave in an atom [1]. The time for one complete revolution of electrons around the nucleus (the period  $T$ ) is

$$T = \frac{2r\pi}{v} = \frac{1}{f}, \quad (69)$$

$f$  is the frequency of rotation. The period of electromagnetic wave,  $T_{em}$ , is

$$T_{em} = \frac{1}{\nu}. \quad (70)$$

The multiplication of (69) with  $\nu$  gives

$$\frac{\nu}{f} = \frac{2r\pi\nu}{v}. \quad (71)$$

Using (60) and (68) from (71) we obtain

$$\frac{\nu}{f} = \frac{|qQ|}{4\epsilon_0 A_0 v}. \quad (72)$$

The electromagnetic wave in an atom can exist as a standing wave [11]. Standing wave does not transmit the energy, but it sways existing energy. If the frequency of the standing wave is  $\nu$ , active power (index  $ap$ ) of the standing wave oscillates with dual frequency  $f_{ap} = 2\nu$ . To electromagnetic standing wave took place in an atom there must be a mutual harmony between two above-mentioned processes. It means that the frequency  $f_{ap}$  of active power must be an integer relationship with the frequency of rotation  $f$ ,

$$f_{ap} = n f, \quad (73)$$

where  $n$  is one of the hole numbers 1, 2, 3, ... Both above-mentioned processes in respect of synchronization are equal, so also applies

$$f = n f_{ap}. \quad (74)$$

Two equations, (73) and (74), can be written in the form of only one expression,

$$f_{ap} = n^{\pm 1} f, \quad (75)$$

or

$$2v = n^{\pm 1} f \quad (76)$$

From (72) and (76) we obtain the velocity of electrons  $v_n$ :

$$v_n = \frac{1}{n^{\pm 1}} \frac{|qQ|}{2\varepsilon_0 A_0}. \quad (77)$$

From (76) follows:

$$v_n = \frac{1}{2} n^{\pm 1} f_n. \quad (78)$$

From (48) and (77) follows

$$s_0 = \sqrt{\frac{Z}{2n^{\pm 1}\beta_n}}, \quad (79)$$

and from (28)

$$\beta_n = \sqrt{1 - \left( \sqrt{1 - \beta_0^2} - \frac{eV}{mc^2} \right)^2}. \quad (80)$$

From (79) and (80) follows

$$s_0 = \sqrt{\frac{Z}{2n^{\pm 1} \sqrt{1 - \left( \sqrt{1 - \beta_0^2} - \frac{eV}{mc^2} \right)^2}}}. \quad (81)$$

The equation (81) is a constant, independent of the variables  $\beta_0$ ,  $n$ ,  $V$  and  $Z$ . For an accurate calculation of the constant  $s_0$  we need only one precise measurement. It can be made through data ionization of hydrogen atoms. In this case in (81) is:  $Z=1$ ,  $\beta_0=0$ ,  $n^{\pm 1}=1$

$$s_0 = \sqrt{\frac{1}{2\sqrt{1 - \left(1 - \frac{eV}{mc^2}\right)^2}}}, \quad (82)$$

with NIST data, <http://www.nist.gov/pml/data/asd.cfm>,  $eV=13.598\ 434\ 005\ 136(12)$  eV and CODATA 2010 recommended values of fundamental physical constants,  $e = 1.602\ 176\ 565(35) \times 10^{-19}$  C,  $m = 9.109\ 382\ 91(40) \times 10^{-31}$  kg,  $c=299\ 792\ 458$  m/s, we get  $s_0 = 8.278\ 691\ 78(11)$ , with relative standard uncertainty  $1.3 \times 10^{-8}$ . This is in accordance with the results obtained in the section III.

## V. THE FINE-STRUCTURE CONSTANT AND THE STRUCTURAL CONSTANT OF ATOM

From (79) we get

$$\beta_n = \frac{1}{n^{\pm 1}} \frac{Z}{2s_0^2}, \quad (83)$$

where  $n^{\pm 1}$  stands for  $n^{\pm 1}=1,2,3, \dots$  or  $n^{-1}=1,2,3, \dots$ , depending on whether the orbits away or closer to the atomic nucleus, respectively.

The maximum amount of  $\beta = v/c$  is 1. This is the case when  $Z$  is a maximum at the same time when  $n^{\pm 1}$  is a minimum (a minimum of  $n^{\pm 1}$  and  $n^{-1}$  is one, i.e.,  $n^{\pm 1}=1$ ). Therefore, the maximum atomic number theoretical is  $2s_0^2$ , and really the maximum atomic number  $Z_{max}$  is an integer of  $2s_0^2$ , i.e.,

$$Z_{max} = \text{Integer} \left( 2s_0^2 \right). \quad (84)$$

We'll connect now three physical quantities, the number of different elements, isotope number and fine-structure constant. The number of different elements is the maximum atomic number  $2s_0^2$ . If we want to describe some isotope (now existing 3 179), it is necessary to define a physical quantity, let's call it 'isotope number' and determine its *unit*. The unit of isotope number can be fraction  $1 / (\text{fine} - \text{structure constant})$  of the maximum atomic number. My proposal is that it's named *boskovic*, in honor Croatian scientists in the 18th century, Roger Joseph Boskovich (Rudjer Josip Boskovic), which dealt with atomistic and described it in its work in Latin *Theoria Philosophiae Naturalis* from 1763. The symbol of this unit can be B.

Then we would have 8 basic units: 1 - meter (m), unit of length, 2 - kilogram (kg), mass units, 3 - second (s), unit of time, 4 - ampere (A), unit of electrical current, 5 - kelvin (K), unit of thermodynamic temperature, 6 - candela (cd), unit of luminous intensity, 7 - mole (mol), unit of amount of

substance, and additionally, 8 - boskovic [B], unit of isotope number, Table I.

We can say now that the *fine-structure constant* is a unit of *number of different elements*. It is the fraction  $1/137.073\,475\,176\,479\,14$  of the *maximum number of different elements*, i.e.,

$$\alpha = \frac{1}{2s_0^2}. \quad (85)$$

Comparing (82) and (85) we get  $\alpha = \sqrt{1 - (1 - eV / mc^2)^2}$ .

Thus,  $S = 1 \frac{0}{7} [B]$  represents the hydrogen  ${}^1_1\text{H}$ ,  $S = 1 \frac{1}{7} [B]$  represents the hydrogen  ${}^2_1\text{H}$ ,  $S = 1 \frac{2}{7} [B]$  represents the hydrogen  ${}^3_1\text{H}$ , ...  $S = 1 \frac{6}{7} [B]$  represents last isotope of hydrogen  ${}^7_1\text{H}$ . Generally, when an atom has  $Z$  protons and  $D$  isotopes, then for his  $i$ -th isotope worth (Fig. 2 and Table I).

$$S = \left( Z + \frac{i-1}{D} \right) [B]. \quad (86)$$

It should be noted that the possible and different formula, say those who would rather  $D$  had assumed the maximum possible number of isotopes of an element, or similar.

#### VI. THE PLANCK'S CONSTANT AND THE STRUCTURAL CONSTANT OF ATOM

The equations (52) and (53) describe the photon energy  $\Delta E_{em} = A\nu$ . According to Einstein's proposal this energy is equal to  $h\nu$ , where  $h$  is Planck's constant and  $\nu$  is the frequency of electromagnetic wave (i.e., the light). This means that  $h = A$  and in accordance with (53)

$$h = A = A_0 + \frac{mc^2 \sqrt{1 - \beta_0^2}}{\nu} - \sqrt{A_0^2 + \left( \frac{m c^2}{\nu} \right)^2}. \quad (87)$$

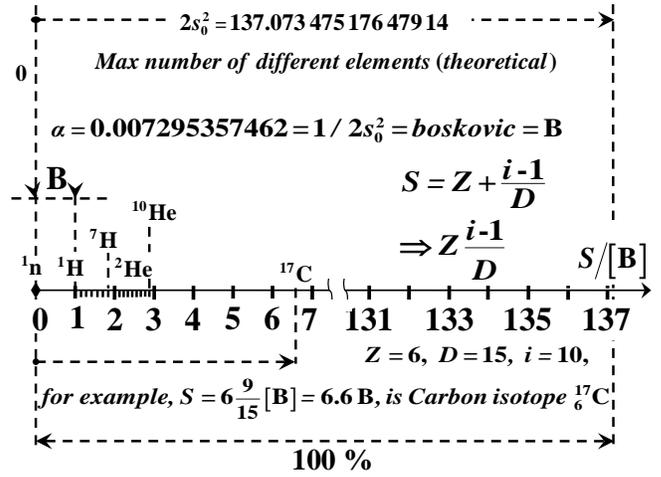


Fig. 2 The fine-structure constant  $\alpha$  presented as a fraction  $1/137.073\,475\,176\,479\,14$  of the maximum number of different elements, and the proposed introduction of unit isotope number named *boskovic* with the symbol  $B$ . In that way one can assign a number  $S$  to each of the isotope (now there are more than 3 179 isotopes)

It is obvious that Planck's  $h$  depends on the frequency  $\nu$ . For low frequencies applies.

$$h \approx A_0 = c \mu_0 e^2 s_0^2. \quad (88)$$

It is more correct to speak of the constants  $A_0$  and the energy of photons

$$\Delta E_{em} = A_0 \nu + mc^2 \sqrt{1 - \beta_0^2} - \sqrt{(A_0 \nu)^2 + (mc^2)^2}, \quad (89)$$

but the Planck's  $h$  and the energy of the photons  $\Delta E_{em} = h\nu$ .

Table I Proposal: the basic units with the addition of unit of isotope number called *boskovic* [B]. The boskovic, unit of isotope number, is the fraction  $1/137.073\,475\,176\,479$  of the maximum number of different elements

| Quantity                  | Sign | Unit     | Symbol |
|---------------------------|------|----------|--------|
| Length                    | $l$  | meter    | m      |
| Mass                      | $m$  | kilogram | kg     |
| Time                      | $t$  | second   | s      |
| Electrical current        | $I$  | ampere   | A      |
| Thermodynamic temperature | $T$  | kelvin   | K      |
| Luminous intensity        | $I$  | candela  | cd     |
| Amount of substance       | $n$  | mole     | mol    |
| Isotope number            | $S$  | boskovic | B      |

Table II An abbreviated list of the fundamental constants of physics and chemistry which are made using structural constant of atoms  $s_0$ , electron mass  $m$ , proton mass  $\mu_p$ , magnetic constant  $\mu_0$  and speed of light in vacuum  $c$

| Quantity                      | Symbol        | Equation                       | Numerical value                     | Unit               | % Difference* |
|-------------------------------|---------------|--------------------------------|-------------------------------------|--------------------|---------------|
| structural constant of atom   | $s_0$         | $\sqrt{\sigma(\chi)Z}$         | 8.278 691 78(11)                    |                    | not known     |
| inverse fine-structure const. | $\alpha^{-1}$ | $2s_0^2$                       | 137.073 475 176 479                 |                    | +0.027        |
| fine-structure constant       | $\alpha$      | $1 / 2s_0^2$                   | $7.295 357 462 212 \times 10^{-3}$  |                    | -0.027        |
| von Klitzing constant         | $R_K$         | $\mu_0 c s_0^2$                | $2.581 280 744 340 \times 10^4$     | $\Omega$           | +0.027        |
| action constant, Planck's $h$ | $A_0$         | $\mu_0 c s_0^2 e^2$            | $6.627 882 090 554 \times 10^{-34}$ | J s                | +0.027        |
| conversion constant           | $K_0$         | $1 / 2\mu_0 c s_0^2 e$         | $1.208 664 052 189 \times 10^{14}$  | Hz V <sup>-1</sup> | not known     |
| ratio $e / h = 1/2 K_J$       | $e / h$       | $1 / \mu_0 c s_0^2 e$          | $2.417 328 104 378 \times 10^{14}$  | A J <sup>-1</sup>  | -0.027        |
| Josephson constant            | $K_J$         | $2 / \mu_0 c s_0^2 e$          | $4.834 656 208 756 \times 10^{14}$  | Hz V <sup>-1</sup> | -0.027        |
| elementary charge             | $e$           | $\sqrt{2\alpha A_0 / \mu_0 c}$ | $1.602 176 565 000 \times 10^{-19}$ | A s                | 0             |
| Rydberg constant              | $R_\infty$    | $m / 8\mu_0 s_0^6 e^2$         | 10 964 733.322 649 5                | m <sup>-1</sup>    | -0.082        |
| Bohr radius                   | $a_0$         | $\mu_0 s_0^4 e^2 / \pi m$      | $5.294 666 854 026 \times 10^{-11}$ | m                  | +0.055        |
| Bohr magneton                 | $\mu_B$       | $\mu_0 c s_0^2 e^3 / 4\pi m$   | $9.276 546 521 260 \times 10^{-24}$ | A m <sup>2</sup>   | +0.027        |
| Nuclear magneton              | $\mu_N$       | $\mu_0 c s_0^2 e^3 / 4\pi m_p$ | $5.052 165 140 176 \times 10^{-27}$ | A m <sup>2</sup>   | +0.027        |

\*Difference value in relation to the Committee on Data for Science and Technology, CODATA 2010

### VII. THE JOSEPHSON CONSTANT AND THE STRUCTURAL CONSTANT OF ATOM

According to (20) and (58) we can write ( $q=e$ )

$$v = \frac{|e \Phi|}{2A_0}. \quad (90)$$

If we divide (90) with the potential  $|\Phi|$ , we obtain the ratio of two constants, which is again a constant. Using (88) we get conversion constant:

$$K_0 = \frac{v}{|\Phi|} = \frac{e}{2A_0} = \frac{e}{2\mu_0 c e^2 s_0^2} = \frac{1}{2\mu_0 c s_0^2 e}. \quad (91)$$

So, (90) we can write using (91):

$$v = K_0 |\Phi|. \quad (92)$$

Equation (92) is reminiscent of Josephson's equation of the inverse AC effect,  $v = (2e/h)U_{DC}$ , where  $2e/h = K_J$  is Josephson's constant  $K_J$ , while  $U_{DC}$  is the voltage at the superconducting junction, analogous to potential  $|\Phi|$  in (90). Using (88) we find

$$K_J = \frac{2e}{A_0} = \frac{2e}{\mu_0 c e^2 s_0^2} = \frac{2}{\mu_0 c s_0^2 e} = 4K_0. \quad (93)$$

### VIII. THE VON KLITZING CONSTANT AND THE STRUCTURAL CONSTANT OF ATOM

If we share the constant  $A_0 = c \mu_0 e^2 s_0^2$  expressed by (48), with  $e^2$ , we get again a constant:

$$R_K = \frac{A_0}{e^2} = \mu_0 c s_0^2. \quad (94)$$

This constant coincides with the von Klitzing's constant  $e^2/h$  which was obtained in the research of the quantum Hall effect, Table II. It should be noted that the achievement of either integer (1, 2, 3, ...) or fractional values (1/3, 2/5, 3/7, 2/3, 3/5, 1/5, 2/9, 3/13, 5/2, 12/5, ...) in this effect is reminiscent of our introduction the numbers  $n^{\pm 1}$ .

### IX. CONCLUSION

The paper describes the determination of the structural constant of atom  $s_0$  and its application in the calculation of physical quantities. This constant is completely determined from Maxwell's equations in the framework of classical physics. Thus defined constant is clear, observable and measurable. This means that this constant is not necessary to further interpret.

The structural constant of atom helps in the interpretation of the fine-structure constant, and it can completely replace the fine-structure constant. The fine-structure constant has been

proposed as a unit of isotope number named *boskovic* [B]. In that way one can classify all the currently known elements and their isotopes, which is now recorded more than 3 200.

In the paper Planck's constant, Josephson's constant, von Klitzing's constant and Rydberg constant theoretically are connected to the structural constant of atom  $s_0$ .

The paper also shows limitations in the application of Planck's constant and proposes a solution for its increased use, through the replacement of Planck  $h$  with  $A_0 = c \mu_0 e^2 s_0^2$  and modification of the expression for the photon energy of  $\Delta E_{em} = h\nu$  to  $\Delta E_{em} = A_0\nu + mc^2 - \sqrt{(A_0\nu)^2 + (mc^2)^2}$ .

This change leads to the correction of Duane-Hunt's law, through which could be carried out verification of the theory presented in this paper at voltages above 20 kV. The verification can also be done by using spectral analysis.

If the presented theory confirmed as correct, it would have a significant impact on the further development of science in general.

#### ACKNOWLEDGMENT

Wolfram Research, Inc. *Mathematica* software is used by courtesy of *Systemcom*, Ltd., Zagreb, Croatia, [www.systemcom.hr](http://www.systemcom.hr). The author thanks Ms. Dubravka Brandic for editing this paper in English, Ms. Srebrenka Ursic, Mr. Damir Vuk and Mr. Branko Balon for the useful discussions.

#### REFERENCES

- [1] M. Perkovac, Quantization in Classical Electrodynamics, *Physics Essays*, 15, 2002, 41-60. Available: <http://connection.ebscohost.com/c/articles/11163931/quantization-classical-electrodynamics>
- [2] M. Perkovac, Absorption and Emission of Radiation by an Atomic Oscillator. *Physics Essays*, 16, 2003, 162-173. Available: [http://www.researchgate.net/publication/229020939\\_Absorption\\_and\\_emission\\_of\\_radiation\\_by\\_an\\_atomic\\_oscillator](http://www.researchgate.net/publication/229020939_Absorption_and_emission_of_radiation_by_an_atomic_oscillator)
- [3] M. Perkovac, Maxwell's Equations as the Basis for Model of Atoms. *Journal of Applied Mathematics and Physics*, 2, 2014, 235-251. Available: <http://dx.doi.org/10.4236/jamp.2014.25029>
- [4] M. Perkovac, Determination of the Structural Constant of the Atom. *Journal of Applied Mathematics and Physics*, 2, 2014, 11-21. Available: <http://dx.doi.org/10.4236/jamp.2014.23002>
- [5] M. Perkovac, Model of an Atom by Analogy with the Transmission Line. *Journal of Modern Physics*, 4, 2013, 899-903. Available: <http://dx.doi.org/10.4236/jmp.2013.47121>
- [6] R. D. Benguria, M. Loss and H. Siedentop, Stability of Atoms and Molecules in an Ultrarelativistic Thomas-Fermi-Weizsäcker Model, 2007, 1-11. <http://dx.doi.org/10.1063/1.2832620>
- [7] D. C. Giancolli, *Physics for Scientists and Engineers* (Prentice Hall, Englewood Cliffs, 1988).
- [8] L. Page and N. I. Adams, *Electrodynamics* (D. Van Nostrand Company, Inc., New York, 1940).
- [9] R. Rüdénberg, R. *Elektrische Schaltvorgänge* (Verlag von Julius Springer, Berlin, 1923).
- [10] M. Perkovac, Statistical Test of Duane-Hunt's Law and Its Comparison with an Alternative Law, 2010. Available: <http://arxiv.org/abs/1010.6083>
- [11] Z. Haznadar and Z. Stih *Elektromagnetizam* (Skolska knjiga, Zagreb, 1997).

# Face-Recognition Based Authentication: Theory and Practice

Thomas Fenzl, Christian Kollmitzer, Stefan Rass, Peter Schartner

*Abstract*—Password challenges are a standard and widespread component in nowadays security systems. Basically any form of authentication at some stage calls for the entry of a particular secret access code. Shoulder-surfing is a laughably simple, yet incredibly powerful way of breaching such security measures. The attack is nothing else than spying on a password from behind the owner, when the access code is entered. Graphical passwords were introduced to thwart this kind of intrusion. Despite many different techniques being around, graphical passwords are difficult to compare to standard password challenges in terms of quality and usability. Taking face-recognition challenges as our case-study, we sketch a framework for measuring the uncertainty (in terms of entropy) that lies within a face-recognition challenge. From the practical point of view, we evaluate the usability of such an authentication system, and report on user’s experiences and their willingness of using such systems in everyday life.

## I. INTRODUCTION

Judging from the good job that cryptographers did over the last decades, an attacker can hardly hope to break a cipher or to forge a digital signature without knowing the secret cryptographic keys. Instead, it is much simpler and reliable to spy on login information, giving access to the required secret keys. A standard password or PIN could be easily spied out, for example, by a quick glance over an owner’s shoulder when the access code is entered, or by listening in to a conversation where the secret key is blabbed to another person. Moreover, social engineering is a collective term for techniques like dumpster diving and others, in which an intruder gathers information about the subject whose identity is to be stolen. In the light of so many people nowadays publishing their most intimate details on social networks, password discovery, thanks to common and frequent password reuse, has become simpler than ever. Insufficient security awareness among users is often a widely open door for hackers, which lets them conveniently circumvent the strong cryptographic protection that may entirely rest on a simple password challenge.

Graphical passwords target at thwarting such trivial attacks as through shoulder-surfing or by eeducing passwords from the user on the phone (via a call from some dubious ”support center”). With the human brain, evolutionary specialized on

recognizing human faces [8, 9], why not use faces instead of alphanumeric symbols to form a pass-code? The idea is not new, and has seen commercial implementation (cf. <http://www.realuser.com/>). Moreover, a variety of interesting alternatives has popped up ever since the idea of graphical passwords has been born (see [10] for a survey).

We wondered why the charming and beautiful idea behind graphical passwords has not received much more attention. Indeed, the sceptic user would for sure like answers to at least these questions:

- 1) Are graphical passwords comparable to standard passwords in terms of security?
- 2) What about usability? Is it easy and feasible to use this alternative?

Giving an affirmative answer to these questions that covers for the entire field of graphical passwords is much beyond the scope of this article. Therefore, we analyzed face-recognition based authentication in terms of theoretical security and practical usability. We shall elaborate on both aspects in the sequel.

### *Face-Recognition Based Authentication*

Similarly as when entering a PIN at an automatic teller machine (ATM), one can equally well challenge the user to recognize a given set of faces. Such a login-screen is displayed in figure 1. The striking difference between such a challenge and a standard password is that a secret key consisting of a sequence of faces is much more difficult to memorize from looking over a user’s shoulder than a PIN or password would be. Consequently, a quick glance from behind a person will be of no substantial value for the attacker, thus rendering shoulder-surfing mostly infeasible. Observe that none of the faces in figure 1 has hair or a background, because both features dramatically simplify distinguishing faces, yet at the same time support restricting attention to few easy-to-tell features. This could defeat the intended resilience against shoulder-surfing or passing on the access code in oral or written form.

This form of authentication raises several questions regarding theoretical as well as practical properties, whereas a simple set of requirements is immediately obvious: first, pictures should never come with a background, for otherwise the user could focus on memorizing particular features, say a tree or an animal, partially visible behind the face. Second, the hairstyle is of significant recall value, as our empirical study in section III indicated, and therefore should not be an attribute making major differences. Third, the faces must be of sufficient distinctness to ease memorizing and recognizing

Institute of Psychology, Universität Klagenfurt, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria, email: [thomas.fenzl@aau.at](mailto:thomas.fenzl@aau.at)

AIT Austrian Institute of Technology GmbH, Quantum Technologies, Department Safety & Security, Lakeside B01A, 9020 Klagenfurt, Austria, email: [christian.kollmitzer@ait.ac.at](mailto:christian.kollmitzer@ait.ac.at)

Institute of Applied Informatics, System Security Group, Universität Klagenfurt, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria, email: [stefan.rass@aau.at](mailto:stefan.rass@aau.at)

Institute of Applied Informatics, System Security Group, Universität Klagenfurt, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria, email: [peter.schartner@aau.at](mailto:peter.schartner@aau.at)

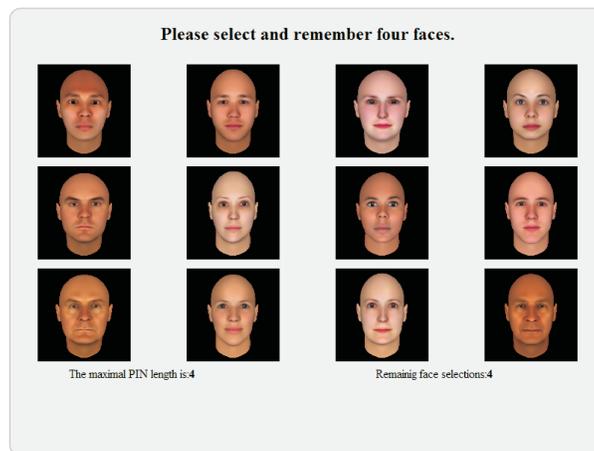


Fig. 1: Face-Recognition challenge for authentication

them later on. Finally, the pool of faces that are presented upon a login challenge *must* be static and fixed, regarding the "personal identification faces" (PIF) as well as the accompanying decoy pictures. Imagine this not being the case, then the adversary may simply record a sequence of login challenges and distill the four or five faces that identically appear in every round. This quickly reveals the access code and defeats the whole system. For similar reasons, the position of the faces (row and column in the login screen) must *not* remain static over repetitions of the login, for otherwise one can just memorize the respective positions to sneak in, and forget about the faces as such.

## II. FACE-RECOGNITION VS. CLASSICAL PASSWORDS – SOME THEORY

The strength of a standard password challenge is quantitatively measured in terms of entropy. Roughly speaking, if a random password  $X$  is chosen, then its Shannon-entropy  $H(X)$  measures the uncertainty that the adversary faces when attempting to guess it. A study from Germany [6] recently came up with the following surprising results: the per-letter entropy for a password was only about 1.9 bit per letter on a word-basis. Even more surprising was the fact that the entropy is hardly increased when the passwords are made longer. This is due to users tending to attach easy-to-guess affixes in order to retain the longer password rememberable. Neither of these phenomena applies to graphical passwords.

Of course, the best case is selecting the password randomly and uniformly from the set of possibilities. This would maximize the password's entropy, but makes memorization much harder. Password policies, such as the prescription to use at least one upper-case, one number and one special character in the password of length at least 6 characters, are there to enforce such a close-to-uniform selection process. Alas, the study of [6] indicated the effect usually not living up to the expectations. Can face-recognition challenges be a remedy? We are not aware of any critical discussion about policies in the context of graphical passwords. Therefore, we shall sketch a few thoughts on this in section II-C.

### A. Dictionary Attacks

Knowing about the usual ways of attacking a password challenge (such as through dictionaries or rainbow-tables), why not set up a dictionary of faces? Social networks can help, since many users maintain huge photo galleries, and finding a picture that served as a mnemonic for a face-recognition challenge appears more than likely. In fact, if we assume a user to have selected a set of, say  $k$ , pictures, how can we reliably remember them? Even if it is not deliberately, a user might tend to choose faces that appear familiar, so as to easily recognize them later on. By that point, we can browse social networks and photo-sharing websites to examine the subject's personal photo gallery and set up a "personalized" dictionary for attacking. The image search facility of many webservices (such as Google, Yahoo, etc.) are vast resources of pictures showing celebrities of all kinds. These may in addition make valuable mnemonics and can be used to set up dictionaries for attacking. Automated software [2, 1] for matching faces against each other is widely available (many digital cameras contain systems that identify faces on a picture for "smart focusing"), hence there appear to be no severe obstacles in running a dictionary attack on a face-recognition challenge, in pretty much the same way as we could do it for a standard password authentication.

### B. Entropy of Face-Recognition Challenges

To get an idea about the extent to which a graphical password challenge is comparable to a standard alphanumeric password, one can compute the Shannon-entropy as a measure of uncertainty for comparison. It turns out that a face-recognition challenge requires a considerably large pool before beating a standard password. For instance, with a selection of 8 out of 64 gives  $\log_2 \binom{64}{8} \approx 32$  Bit of entropy at best (assuming that the order of images does not matter, and the choices are uniform). Things look much better when we compare it to a PIN-challenge, were we have a selection of 5 out of 25 images roughly comparable to a 5 digit PIN-code challenge. Still, from an information-theoretic perspective, face-recognition challenges disregarding the ordering are almost equally powerful as many standard alphanumeric

password challenges, and notably, are much more resilient against shoulder-surfing or key-logging.

### C. Policies

Normally, a password policy serves two purposes:

- 1) Prescribing features of the password, such the guideline to include at least one character, special symbol, digit, etc., prevents the user from choosing "easy-to-guess" passwords (e.g. names, dates, words from his native language or similar). Ideally, the policy enforces a choice close to the uniform distribution, hence maximizing the entropy (uncertainty for the adversary) when choosing the password.
- 2) Expiration dates for passwords are necessary to reduce the chances of espionage through shoulder-surfing or general eavesdropping.

While the above requirements obviously strengthen the security of alphanumeric passwords, it is worth discussing whether or not such policies are needed for graphical passwords as well. There are three points at which the attacker can attempt eavesdropping on a password: the terminal where it is entered, the channel over which it is sent, and the receiver who grants or denies access. While espionage in the latter two cases can be impeded by standard cryptographic techniques, the first scenario is worth a closer look. Notice that much unlike alphanumeric passwords, key-loggers are no direct threat to a graphical password. Moreover, since face-recognition based authentication is *designed* to thwart spying on the pass-code, the second of the above arguments can be put to question. In general, the quality of a graphical password is much harder to measure accurately than for an alphanumeric password, and preventing the user to choose from a pool of easy-to-guess faces is possible by excluding such faces from the pool in the first place. Much work is to be done before an affirmative answer about policies for graphical passwords can be given, yet the problem appears interesting and nontrivial.

### D. Psychological Aspects of Face-Recognition

The human brain is evolutionary specialized on recognizing and memorizing natural and schematic faces [9, 8] and research in neuroscience has provided significant evidence for the existence of a specific cerebral area that is responsible for facial recognition (see e.g. [3]). While a glance at a face may reveal an enormous amount of information, including a person's age, gender, emotional state and others, the process of face recognition usually starts with detecting the eyes, matching them with the eye prototype and the eye-eyebrow template (templates are characterized by distances between certain regions in a face, e.g. the distance from the chin to the eyes or the distance of the eyes and their displacement to the eyebrows), continues with feature-based detection of the nose and mouth and then more and more combined pattern-matching processes occur [4].

However, when developing face-recognition based authentication systems, not only physiognomic aspects are of particular interest but memory psychology and cognitive psychology have to be considered as well. In particular the

processing capacity of the human brain will be one of the limiting factors for the complexity and maximum length of the graphical password. Previous research shows that the short-term memory is limited to the processing of  $7 \pm 2$  chunks, that is, the specific units to be recalled, which are determined by the structure of the presented material. For example, individuals may usually recall six monosyllabic words – where each monosyllabic word represents one such chunk – or three words with four syllables (each) in reverse order. However, they will fail in doing so when they are presented nine monosyllabic words or six words with four syllables (each), as the number of units to be recalled in reverse order exceeds the maximum number of chunks that may be processed by the short-term memory. Regarding face-recognition, and face-recognition based authentication in particular, the crucial question is, how these chunks are determined in the process of memorizing a sequence of faces. Certainly it is not trivial to answer this question, as the mental representation of a face is not necessarily comprised of particular characteristics and determinants of the face, but may also be reflected by an overall facial expression [11]. Hence, giving an affirmative answer to this question is way beyond the scope of this paper and requires a series of experimental investigations.

## III. PRACTICAL EXPERIENCE

We analyzed the practical usability of a prototype for face-recognition based authentication in an empirical field study, applying a mixed-methods design [5] to ensure a comprehensive understanding of the processes related to memorizing and recalling sequences of faces.

### A. Psychological test scenario and sample

After a brief introduction participants in the study were randomly assigned to one of two groups and had to register in the system with their socio-demographic data (age, sex and ethnicity), a pseudonym and a graphical pass-code. In particular they had to select and memorize a sequence of either three (group A) or four (group B) faces from the login-screen (see figure 1), which depicted twelve randomly assigned and synthetically generated faces. Double-selection of faces or taking back previous choices was prohibited in the registration phase. Registration was successfully completed upon confirmation of the PIF (similarly to a standard password choice and re-entry query). Furthermore participants were asked to authenticate in the system after some waiting period, where group A had to recall the three faces of their pass-code (PIF) in order. Group B, which was initially assigned a code-length of four faces, was randomly separated into two distinct groups, of which group B1 had to recall the faces in order and group B2 without order. Again, the PIF had to be selected from a login screen (see figure 1) showing the twelve exact same but differently arranged faces as during the registration. The authentication process was repeated either until successful authentication or until termination by the user.

While the field study was conducted, the prototype registered the duration of the registration, waiting and authentication period as well as the error rates in a MySQL-database.

After completion of the authentication process participants were asked to take part in a short guideline-based interview. In the interviews qualitative data on the processes related to selecting, memorizing and recalling the PIF as well as on usability and subjective complexity were collected.

Our sample consisted of 189 successfully registered participants of which 57 took part in the interview-survey. The participants had an average age of  $\bar{x} = 23.4$  years (ranging from under 15 to over 60 years) and all groups showed a homogenous distributions concerning sex. Quantitative data was analyzed using inferential statistics and qualitative content analysis [7], in particular the procedure of inductive category formation, was applied to explore the interviews.

## B. Results

To begin with, the majority of respondents judged the idea of face-recognition based authentication rather positive (75%) and indicated a very good usability of the prototype in the interviews. Only less than one fifth of the respondents stated that they would prefer alphanumeric access-codes rather than graphical passwords. In evaluating the useability of the system, we report the average times ( $\bar{t}$ ) for registration and authentication, along with their respective standard deviations ( $s$ ) as a measure of variability.

### Quantitative Analysis

The statistical analysis yielded firstly that the average duration of the registration process was moderate  $\bar{t}_{\text{reg}} = 47.6$  sec (standard deviation  $s_{\text{reg}} = 17.2$  sec) and increased significantly (two-sided independent two-sample t-Test,  $p = 0.004$ ,  $\alpha = 5\%$ ) with the length of the PIF. On average it took the 93 participants of group A about  $\bar{t}_{\text{reg}} = 44$  sec ( $s_{\text{reg,A}} = 14.8$  sec) to select and memorize a sequence of three faces, while it took the 96 participants of group B, whose code consisted of four faces,  $\bar{t}_{\text{reg,B}} = 51$  sec ( $s_{\text{reg,B}} = 18.7$  sec) on average. Remember that both groups had to memorize the order of faces for the confirmation of their code in the registration phase.

Of the successfully registered people, 43% authenticated successfully in the system in the first attempt. Another 27% completed the authentication process after several more attempts and 30% either did not show up for authentication or terminated the process. However, we were unable to distinguish between the two latter groups based on our available data. Although we may state that participants, who failed to successfully complete the authentication process, obviously had trouble with recalling their PIF correctly, we can only speculate on the motivations of registered users for choosing not to show up for authentication.

Secondly, the statistical analysis revealed that participants were able to recall their PIFs rather quickly  $\bar{t}_{\text{auth}} = 17.2$  sec ( $s_{\text{auth}} = 7.7$  sec) and users, who had to remember the faces in order, entered their PIFs faster than those for whom the order was optional in the authentication process. Statistically speaking, the average authentication times for the participants of group A, who remembered their three faces in order in  $\bar{t}_{\text{auth,A}} = 14$  sec ( $s_{\text{auth,A}} = 4.6$  sec), and for the participants of

group B1, who were able to recall their four faces in order in  $\bar{t}_{\text{reg,B}} = 16.9$  sec ( $s_{\text{reg,B}} = 6.3$  sec), were of equal length. By contrast, with  $\bar{t}_{\text{reg,B}} = 22.9$  sec ( $s_{\text{reg,B}} = 9.5$  sec) it took the participants of group B2 significantly longer to enter their four faces with the order being an option (a two-sided Mann-Whitney-U-Test came back with a  $p$ -value of  $p = 0.0001$ , tolerating a false-negative rate of  $\alpha = 5\%$ ).

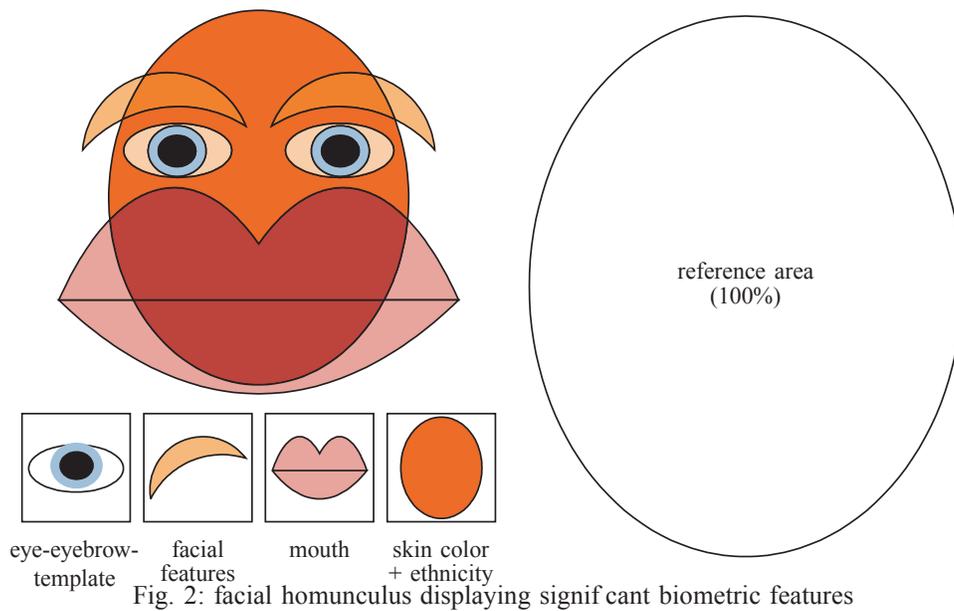
While the majority of participants (75%) estimated the complexity of the face-recognition based registration and authentication rather easy in general, half of the respondents in the interviews stated that memorizing the graphical password (PIF) was more difficult than recalling it and 39.1% stated just the opposite. A fraction of 10.9% rated memorizing and recalling equally complex.

At first sight this result seems surprising, particularly when assuming that the same cognitive processes are at work during face-recognition in the registration and authentication phase. However, a closer look also reveals a considerable difference between the two tasks of memorizing and recalling the PIF in our setting, which originates from the fact that users are presented exactly the same twelve faces – although in different arrangement – on the login screen during registration and authentication. With this knowledge, which was given to all participants in the introduction, a prudent user should firstly get an overview of all the faces on the login screen and potentially even memorize all of them. In a second step, he would then pick the required number of faces for his PIF, while focusing on how he may distinguish these faces from the other faces in the pool. In doing so, the user simplifies the process of face-recognition during authentication. In order to successfully recall the PIF, he just has to remember the distinctive features (which may originate from biometric and emotional determinants as well as from an overall facial expression) and perhaps a correct order.

It is very likely that the users, who stated that memorizing the code was more difficult than recalling it, pursued such a strategy. Hence, a possible explanation for the observed phenomenon, namely that memorizing the graphical password (PIF) was more difficult than recalling it for some users and easier for others, may be an insufficient memorization of the PIF during the registration in the second group, which may have caused hesitance during authentication. In particular, the participants who stated that recalling the code was more difficult than memorizing it, might have focused too much on the faces which constituted their code while not paying enough attention to the distinction against the other faces on the login screen. In addition an increased familiarity with the system during authentication – after already having dealt with face-recognition based authentication during registration – may have played a role for participants who found recalling the code to be easier than memorizing it.

### Qualitative Analysis

The qualitative content analysis of the interviews aimed at condensing the determinants of the processes related to selecting, memorizing and recalling the PIF in the face-recognition based authentication. We found that in addition



to the overall impression of the faces (physiognomy, facial expression, etc.) three fourths of the respondents considered up to three biometric characteristics or a succession of these determinants within the code. The distribution of the number of considered biometric features was roughly as follows: 18% considered one feature, 26% two features, 30% three, 16% four and 11% considered five or more features. The detailed analysis of the data showed that respondents paid particular attention to *skin color* and *feature-based detection of the mouth* (e.g. open or closed, wide or narrow, etc.). The *eye-eyebrow template*, the perceived ethnicity of the persons represented by their faces, *facial features* as well as "other" determinants such as the opulence of the face (slim or tubby), the perceived sex, etc. played a secondary role.

To proportion the considered biometric determinants and their relevance for face-recognition based authentication in relation to the size of a face, a "face-homunculus" may be drawn (see figure 2). The reference area, in this case an oval, reflects 100% of the allocable space, which represents the size of the entire face. The fraction of the area covered by a single biometric characteristic then reflects the relative frequency of this determinant in the data. In particular, the item *mouth* (short for *feature-based detection of the mouth*) was relevant in 32% of the cases and thus covers 32% of the total reference area (this includes the parts overlapping the skin in the abstract picture). Similarly *facial features* as well as the *eye-eyebrow-template* each cover 5% of the reference area, as they were found to be relevant in 5% of the cases. Notice that the size of each component (feature) is to be taken relative to the reference oval, and not relative to each other. For illustration purposes, the relative frequencies of the two determinants *skin color* (relevant in 43% of the cases) and *ethnicity* (relevant in 5% of the cases) are displayed as one combined feature, named "skin color", which then covers 48% of the entire reference area. The pool of "other" determinants (remaining 10%) has not been illustrated in figure 2.

Moreover, our work showed that emotional aspects played an important role when memorizing and recalling the PIF. These aspects can be subdivided into three clusters: Firstly, a facial expression perceived by 45% of the subjects. The importance of this phenomenon has already been highlighted above. Secondly, there are basic emotions such as joy (21% relative frequency), anger (12%) and others (e.g. fear, sadness, etc. totaling to 2%), which people believed to recognize in the faces. And finally, emotions evoked within participants when viewing the faces on the login-screen, were relevant in another 10% of the cases.

In line with the above stated arguments regarding the diametrically opposed results on the perceived complexity of memorizing and recalling the PIF, we also found that participants suggested adding different hairstyles, beards, a greater variety of skin colors and other features like tattoos or make-up to the faces, in order to simplify distinguishing the faces on the login screen and the memorization of the code. Of course, adding such characteristics is not suitable for the practical implementation of face-recognition based authentication. Although such measures might improve usability by significantly increasing the recall value of faces, they at the same time encourage restricting attention to few easy-to-tell features and facilitate the use of mnemonics, thus defeating the intended resilience against shoulder-surfing and passing on the access code in oral or written form.

#### IV. CONCLUSION

*a) Lessons learned from the Field-Trial:* Based on the insights gained from our field trial we may summarize that, regarding cognitive performance, the most consistent effects in face-recognition based authentication were observed with PIFs consisting of three faces, regardless of the complexity of the faces. Moreover our findings suggest that giving the PIF a structure, that is, asking respondents to recall the faces in order, *supports* memorization of the access-code. This effect is doubly positive, since it additionally increases the entropy.

However, further research is to be undertaken to corroborate this hypothesis. Additionally the practical usability of face-recognition based authentication has been rated very positive in the interviews, thus making it an attractive option that should be paid more attention by researchers and practitioners.

b) *Outlook*: From a theoretical point of view, face-recognition based authentication can be analyzed by simple means of combinatorics and information-theory. Our numerical analysis indicated the system comparable to standard password or PIN challenges. However, with the application of PIFs, the threat of shoulder-surfing may be dramatically weakened. Since faces can be generated synthetically using a small number of parameters (storable on a smartcard), even a hypothetical global roll-out is at least technically feasible. Given the technological feasibility and people's observed openness towards this type of authentication, we are curious to see whether users will go on facing PINs or whether they will soon be faced by their personal identification faces.

#### REFERENCES

- [1] K. An, D. Yoo, and M. Chung. An efficient fully automatic face tracking using binary template matching. In *Proceedings of The Ninth International Symposium on Artificial Life and Robotics*, pages 37–40, Beppu, Japan, Jan. 28–30 2004.
- [2] E. Arnaud, B. Fauvet, E. Memin, and P. Bouthemy. A robust and automatic face tracker dedicated to broadcast videos. In *IEEE international conference on image processing*, Genes Italy, 2005.
- [3] L. Betts and H. Wilsons. Heterogeneous structure in face-selective human occipito-temporal cortex. *Journal of Cognitive Neuroscience*, 22(10):2276–2288, 2010.
- [4] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.
- [5] B. Johnson, A. Onwuegbuzie, and L. Turner. Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1:112–133, 2007.
- [6] T. Maus. Das Passwort ist tot – lang lebe das Passwort! *DuD – Datenschutz und Datensicherheit*, 8:537–542, 2008.
- [7] P. Mayring. On generalization in qualitatively oriented research. *Forum: Qualitative Social Research FQS*, 8(3):Art. 26, 2007.
- [8] O. Pascalis and D. Kelly. The origins of face processing in humans: Phylogeny and ontogeny. *Perspectives on Psychological Science*, 4:200–209, 2009.
- [9] C. Shannon and W. Weaver. *A Mathematical Model of Communication*. University of Illinois Press, 1949.
- [10] X. Suo, Y. Zhu, and G. S. Owen. Graphical passwords: a survey. In *21st Annual Conference on Computer Security Applications*, page 10 pp., 2005.
- [11] G. Van Belle, P. Graef, K. Verfaillie, T. Busigny, and B. Rossion. Whole not hole: Expert face recognition requires holistic perception. *Neuropsychologia*, 48:2620–2629, 2010.

# Co-simulation of Redundant and Heterogeneous Modelling Scales for a Phenomenological Approach

Sébastien Le Yaouanq, Christophe Le Gal, Pascal Redou, and Jacques Tisseau

**Abstract**—Complex systems were first modelled by means of differential equations. Next, multi-agent methods tried to focus on the elements therein but were limited, amongst other things, by computation capabilities. Nowadays, more and more works suggest to come back to the original point of view and to adopt a phenomenological approach to study interactions which form their dynamics, while keeping multi-agent systems assets. Multi-interaction systems allow the construction of efficient models from a macroscopic point of view as a superposition of phenomena. A drawback is that we often have to set empirical parameters in these descriptive models. Moreover, there can be no assurance that a priori chosen values for these parameters stay valid throughout the simulation due to the dynamics of the system. To respond to this problem, we expose in this paper a redundant multiscale architecture which is based upon the fact that we can establish models of a same phenomenon at heterogeneous time and space scales. Parameters of a macroscopic model are in fact related to the dynamics of the system at the microscopic scale. Inspired by the Heterogeneous Multiscale Methods and co-simulation, we present a software architecture to perform redundant simulations of a system at different levels of description. We also show how this architecture can be used to operate auto-adaptive simulations of phenomenological models, without any additional cost, by taking advantage of unused processor cores. Then we illustrate our architecture with a new tool for offshore structures design optimization.

**Keywords**—Complex systems, multi-agent systems, multiscale methods, phenomenological approach, co-simulation, ice modelling, offshore structures, design optimisation.

## I. INTRODUCTION

THERE are usually two opposite points of view for the modelling of complex systems. First, we can choose to represent individually the entities of the system, by agents for example [1]. In general, their great number is a major obstacle both to simulate the model and to identify global behaviors.

The second way is to focus on the interactions which induce the dynamics of the system. This phenomenological approach consists in modelling phenomena on the basis of the observations or experimental results at our disposal. This modelling method is particularly relevant in an industrial context. As a matter of fact, we often have to deal with empirical laws from domain experts who are faced with the reality in the field, since theoretical models are unavailable.

For that purpose we use multi-interaction systems that are a declination of classical multi-agent systems. Agents do not represent the entities anymore but they embody the phenomena as a macroscopic description, generally with the help of ordinary differential equations. As a result of this abstraction, the computation time can be significantly reduced. This method has been repeatedly and successfully applied [2] [3] [4], validated [5] [6] and today, there are several models and methodologies that can be used to experiment complex systems with multi-agent systems [7] [8]. The price to pay is that we often have to define parameters for macroscopic laws in each interaction-agent [9]. Assuming we could set values *a priori*, some of these parameters can however fluctuate during the simulation because of the dynamics of the system.

We address this problem by considering that the parameters of a macroscopic model are in fact often related to the dynamics of the system at the microscopic scale. For instance, the diffusion phenomenon expressed by the Fick's laws rests upon a diffusion rate that averages the Brownian motion of particles [10]. Thus we propose to get the most of the knowledge we have about the phenomena, by implementing redundant simulations of a system at different description levels. We discuss in the next section a multi-model software architecture which allows online parameters evaluation, so as to perform co-simulation.

We will also show how this strategy can be extended to benefit from multi-core processor. Multi-agent simulations are not easily parallelizable because of the agents interconnections. Thus we suggest to use the available computation units to perform an implicit parameters optimisation by means of parallel simulations of a same model. The goal is to operate auto-adaptive simulations of phenomenological models, without additional computation time.

Beyond the classical study cases, complex systems modelling and simulation have become a key issue in many industrial processes, such as designing and prototyping. We illustrate in the last section the implementation of our architecture with a new simulation tool for the design of offshore structures as a real case.

## II. COUPLING DESCRIPTION LEVELS

In the context of a phenomenological approach, we are interested in modelling complex systems as a superposition of autonomous phenomena. Thus we focus on a macroscopical and global description of the dynamics of the system. It is obvious that this way the accuracy is abandoned in favour of the efficiency. This approach is very successful for many

S. Le Yaouanq, P. Redou and J.Tisseau are with Lab-STICC UMR6285, UBO / ENIB European Center for Virtual Reality Technopole Brest-Iroise, 29280 Plouzané, France email: leyaouanq@cervval.com

S. Le Yaouanq and Christophe Le Gal are with Cervval, Brittany, France (<http://www.cervval.com>)

problems but it implies introducing empirical closures and parameters in equations. Besides, by definition of a complex system, these parameters are most of the time related to the dynamics of the system which is unpredictable. Therefore it seems to be necessary to use new modelling tools to allow an automatic recalibration of our macroscopical models, by means of multiscale methods.

#### A. Multiscale methods

Multiscale methods are usually splitted into two categories. The first one aims at using different time and/or spatial scales into a global simulation. These include adaptive mesh refinement methods [11] and multigrid methods [12]. They offer a better accuracy but their computational cost remains high. Indeed, their efficiency closely depends on the smallest time step used in the simulation. We would like to avoid this problem by keeping a macroscopic description level for the whole model.

Thus, our approach looks more like the second category of multiscale methods, among which stand equation-free methods [13], patch dynamics methods [14], quasi-continuum methods [15] and heterogeneous multiscale methods (HMM) [16]. Their objective is to couple redundant models which describe the same part of the system at different scales. These models are simulated separately and micro-scale results are used to feed the macro-scale simulation. In some particular cases, the macroscopic model is not explicitly available or becomes invalid. We use therefore a microscopic model to supply the necessary data for the macroscopic model.

Let us consider a macro-scale with a state variable  $U$ . We have seen that this state depends on the dynamics of the system and/or parameters. We have at our disposal a microscopic model that describes the microscopic state variable  $u$  of the same system. The two scales are related one to each other by the use of reconstruction ( $R$ ) and compression ( $Q$ ) operators:

$$\begin{cases} Q.u & = U \\ R.U & = u \end{cases} \quad (1)$$

with the property  $R.Q = I$ , where  $I$  is the identity operator. The role of these operators is to translate the system structure and dynamics from a scale to the other. The general idea is indeed to make round-trips between the two scales, as illustrates Figure 1. Heterogeneous multiscale methods suggest a top-down approach: as soon as there is a lack of data in the macro-scale, we use  $R$  to rebuild a micro-scale simulation from the macro-scale state. This reconstructed microscopical state is simulated over a short duration. Thus we can make data estimations with the help of  $Q$  so as to set new parameters in the macro-scale. The main difficulty lies in the definition of these operators. In the absence of any formal method, we assume that the translations between the two description levels may be necessarily guided by domain expertise [17].

The HMM framework has proved to be very useful in guiding the design of complex systems simulation. It helps to transform multiscale modelling from a somewhat ad hoc practice to a systematic technique. However, it suffers from a lack of general implementation for now. Thus we propose



Fig. 1. Schematics of HMM general framework [16]. Two modelling scales are coupled by means of translation operators. The compression operator defines rules to refine the description and the reconstruction operator averages micro-scale data to feed the macro-scale.

to implement it in the context of multi-agent simulations, by means of co-simulation.

#### B. Co-simulation architecture

Our will is to build a software architecture which tends to generalise the redundant simulation of heterogeneous modelling scales. In this regard, we have to develop tools for each step of the method we described above. First, we need to detect the macro-scale defaults that require more precise simulations. As said previously, they can appear further to a structural evolution induced by the dynamics of the system or an intervention of the user. Thus, we propose to add a watchdog mechanism to the macro-scale in the form of an autonomous agent. This agent has a global view of the execution of the model and structure. It is able to capture abnormal numerical variations and can also have experts knowledge and empirical rules on board, in order to build requests for recalibration. A request relates to a unique parameter and takes the form of a snapshot of the macroscopical model which serves as a base for the consultation of the microscopic model.

The difference of abstraction levels in microscopical and macroscopical models often means making joint use of multiple modelling paradigm. It could be possible to formulate all the models in the same formalism [18]. But, from a modularity perspective, we favour a co-simulation approach to implement our redundant simulations architecture. The co-simulation is a design technique commonly used in the field of electronics which reminds the multi-agent concept. In this context, we consider a set of subsystems that are simulated in a black-box and distributed manner, in order to solve a coupled problem. In our case, each model can be implemented in its own formalism. Hence, we can use existing models, thus limiting any specific developments apart from inputs/outputs managers.

Whether in the field of electronics or computer science, a distributed approach requires the subsystems to be connected one to each other. In this way, they are able to exchange data during the simulation. This is achieved *via* a co-simulation bus. The bus carries out the requests formulated by the watchdog-agent of the macroscopical model to all the other connected modules which may be of three different types: an auxiliary

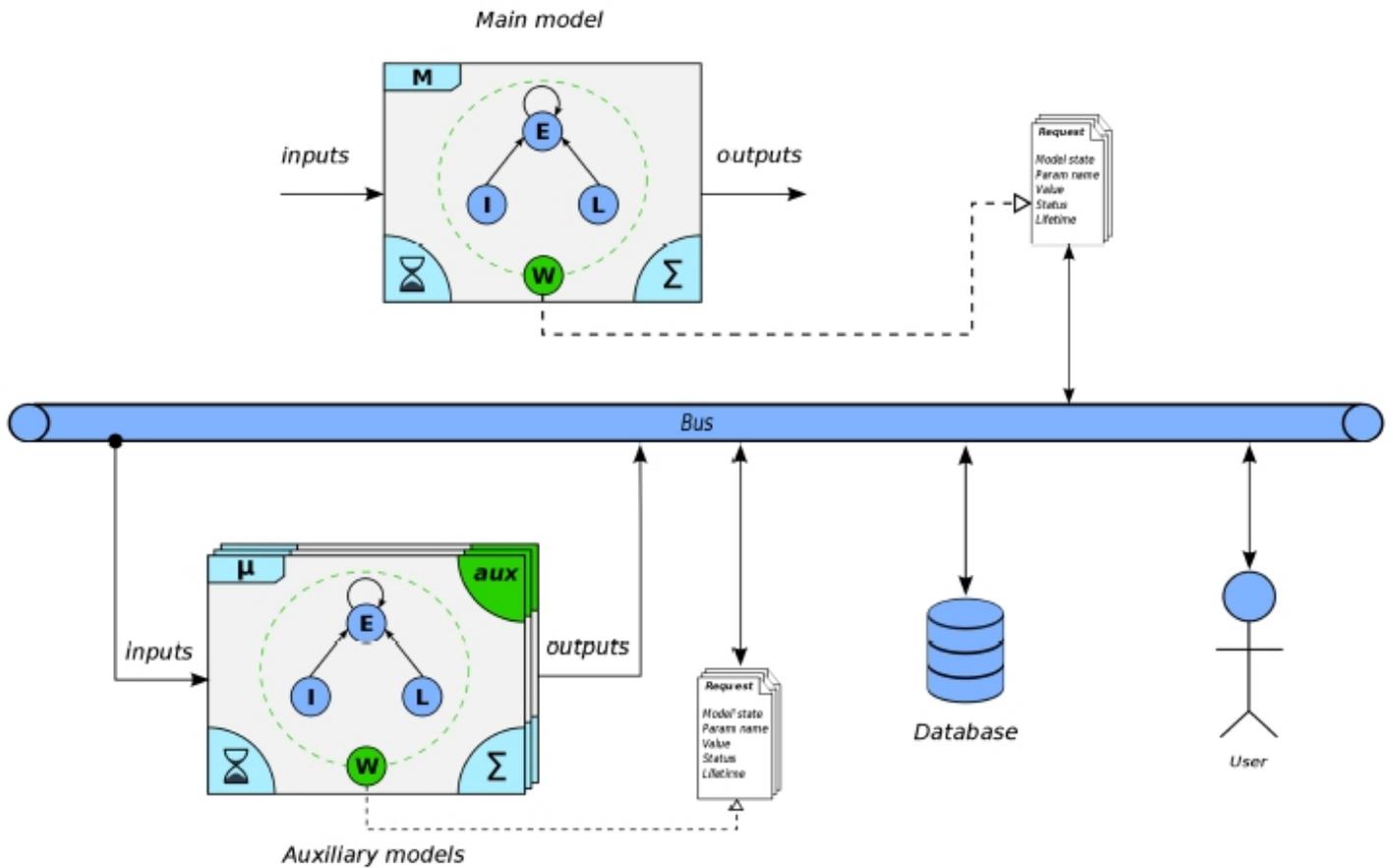


Fig. 2. Co-simulation architecture. We consider the macro-scale as the main model. It contains a watchdog-agent that controls its structure and numerical variations. When this agent detects a default, it builds a request with the macro-scale state and pushes it to the co-simulation bus while the macro-scale is still simulating. The request is thus transmitted to the auxiliary models, databases and users connected to the bus. They all check their own ability to answer the request. In the case of an auxiliary model, the process can be recursively performed. They may or not fulfill the request that is returned to the watchdog-agent which applies the result on the main model.

model, a database or a user. Each of them checks its ability to answer the request according to the data the request contains, i.e. the macro-scale state is sufficient to set initial conditions and that the requested parameter is listed as one of their outputs. Then databases are searched and users are prompted to enter a valid value. In the case of an auxiliary model, we have to translate the macro-scale state so that it corresponds to the model inputs. This is the reconstruction operation of the HMM methodology. As said previously, we use domain expertise to facilitate the procedure. Thus, we assume that the auxiliary models are selected *a priori* knowing the macroscopical model weaknesses. Thereby, we can build meaningful links between macro-scale and micro-scale variables.

Once the auxiliary model is initialized, we have to take account of several technical constraints. We wish to keep real-time or almost real-time interactivity with the macro-scale simulation. This is one of the main focuses of our phenomenological approach. With this aim in mind, we do not pause the simulation. We must therefore ensure that the request is answered in an acceptable time compared to the duration of macro-scale computation steps. Otherwise, the calculated value of the parameter would be out of kilter with the new macro-scale state. This implies that we have to

simulate the auxiliary model over a limited time. However, we can not expect a correct result unless we leave the model enough time to reach a steady-state conducive to measures and extrapolations. We need to come to an agreement on both the efficiency and accuracy. The requests are thus fulfilled and returned to the macro-scale. The watchdog-agent monitors its requests and applies the new parameters values on the macro-scale interaction-agents.

Figure 2 gives a global view of our architecture. Models are represented as boxes containing interaction-agents acting on entities. More details about their construction can be found in [9]. Even if the auxiliary models do not need to be interaction-agents based models, we've decided to highlight here our architecture's recursivity. Indeed, if an auxiliary model requires some calibration during the simulation, its watchdog-agent can in turn send requests to the co-simulation bus.

We will now seek to adapt this first implementation to take advantage of multi-core processors and optimize the computation time of expensive models simulations.

### III. CO-SIMULATION STRATEGIES

Ideally, the micro-level simulation is able to explicitly compute macro-level parameters. In such a case, the strategy

is quite straightforward: the micro-level simulator computes parameters which are then used by the macro-level simulator. Most of the time however, no such explicit computation is possible, either because the macro-level parameters are meaningless in the micro-level, or because they are not an output of the micro-level simulation. For example, in a particle-level simulation, a temperature has no direct meaning; temperature is neither an output, nor a parameter of the micro-level simulation. Such a particle centered simulation could not be used to explicitly compute the temperature parameter of the macro-level simulation.

It may be possible to explicitly compute some parameters but not all of them. In practical cases, in complex systems, all parameters are dependent, which makes impossible to compute some parameters independently of the others. We must therefore consider that the micro-level simulator is only able to output validation results for macro-level.

Therefore, the idea is to perform an implicit computation of the parameters set: micro-level runs in parallel with one or many macro-level simulators, using one or many parameters set. As soon as the micro-level simulator has produced a result for a given time interval, the macro-level simulation whose results are the closest to the micro-level is selected. A new series of parameters set is chosen by an optimization algorithm and the operation is repeated.

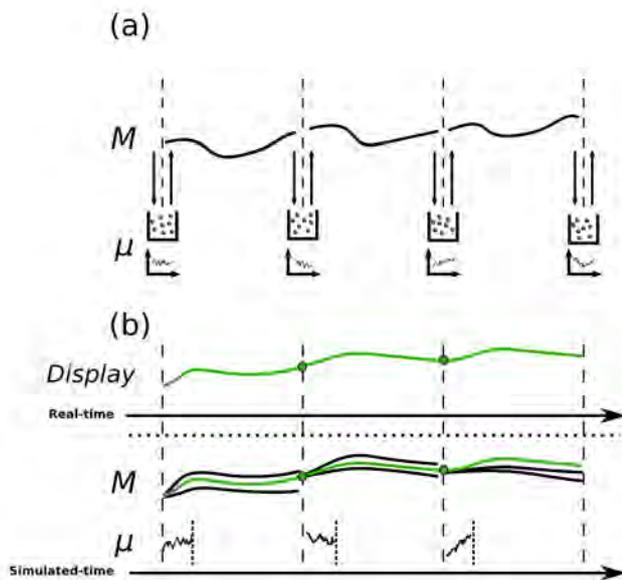


Fig. 3. Comparison of two co-simulation strategies. (a) In the first case, the micro-scale ( $\mu$ ) is used as a parameter estimator. Its initial state is computed from a snapshot of the macro-scale ( $M$ ). The micro-scale simulation is performed for a short duration and data are averaged to calibrate the macro-scale. (b) Three different macro-scale simulations are run with distinct parameters sets ( $M$ ). Meanwhile, a micro-scale simulation is also performed ( $\mu$ ), on a shorter simulated time interval. When the micro-scale simulation is completed, its results are compared to the one, at the same simulated time, of the three macro-scale simulations. The best macro-scale simulation is, *a posteriori*, selected for display on the interactive software interface (*Display*).

To ensure continuity of the displayed results, a small time shift exists between the simulated time and the results used by the interactive software, in order that only macro-level simulators using already selected parameters set are shown to the user.

Of course, we could refine the computation of the parameters by interpolating between best performing parameters set; but that is a matter of optimisation algorithm, which is not in the scope of this paper.

#### IV. APPLICATION TO OFFSHORE STRUCTURES DESIGN

##### A. Context

Designing of offshore structures for Arctic conditions is accompanied by number of challenges related to complex phenomenon of ice failure and its interaction with rigid bodies. The phenomenon of ice interaction with offshore structures of different shapes was investigated by many scientists using analytical and empirical techniques [19] [20], and numerical methods [21] [22]. These methods are widely used for prediction of ice behavior and ice loads exerted on offshore structures operating in ice-infested waters. Nevertheless, the prediction of ice loads is still a challenge, as ice may fail in different mode or in combination of modes depending on ice type, ice thickness, ice mechanical properties, structure geometry, interaction velocity, etc.

The need for an ice simulation tool for offshore platforms design is twofold: the first is to simulate the flow of ice around a fixed or floating platform structure to ensure there is no excessive pile-up and encroachment on the topside facilities and the second is to predict the loadings on the structure so they can be minimized by design and to check they are consistent with the relevant codes and standards. Given the economic and environmental challenges of Arctic developments, any design optimization that minimizes cost and enhances safety is seen as vital.

In June 2012, Technip signed an agreement with Cervral (a specialist software company in Brittany, France) and Bureau Veritas to develop an ice-modelling simulation program called Ice-MAS (Ice Multi-Agent Simulator) (see Figure 4). The original approach of this tool consists in modelling of reciprocal actions among the objects involved into interaction, such as: offshore structure, ice sheet/ice floes, ice blocks due to ice sheet or ice floes failure, water, currents and seabed. The interaction among the objects is simulated by physical models selected according to interaction scenario, shape of structure and environmental conditions. The interaction process is presented in Figure 5, where each arrow represents a simulated phenomenon.

The purpose of this article is not to detail the simulator implementation. Interested readers may refer to the dedicated paper [23]. We would only note that the results of the software validation show a good agreement with ice basin tests and full-scale measurements.

We are now interested in optimizing the computation time for long-term simulations. Indeed, an increase of simulated duration is accompanied with an explosion of ice fragments number. In this context, it becomes more and more difficult



Fig. 4. Ice-loading process in front of a conical structure. The need for an ice simulation tool for offshore platforms design is twofold: the first aim is to simulate the flow of ice around a fixed or floating platform structure to ensure there is no excessive pile-up and encroachment on the topside facilities and the second is to predict the loadings on the structure, that so they can be minimized by design, and to check they are consistent with the relevant codes and standards.

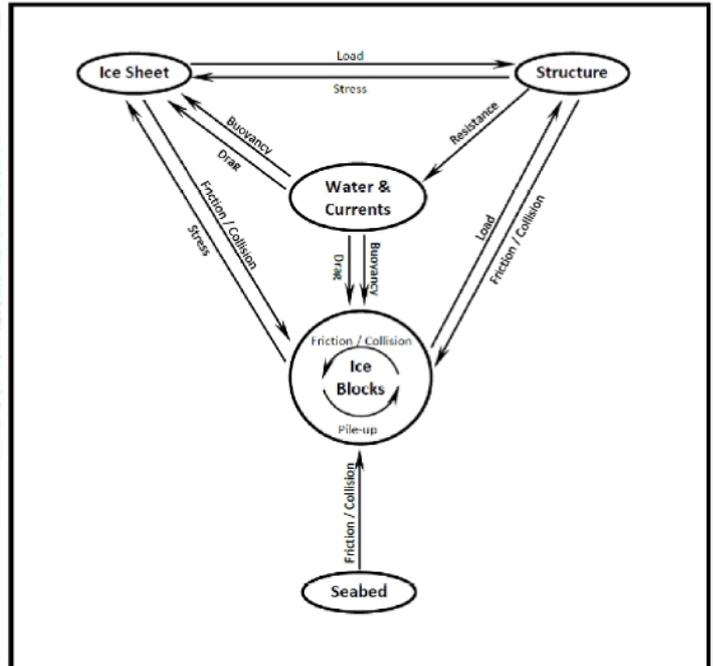


Fig. 5. Ice-MAS interaction diagram. The program is based on a multi-model approach to predict ice behaviour. Accounting for water and mutual reciprocal actions among offshore structure, ice sheet/floes, ice blocks due to ice sheet or ice floes failure, water, currents and seabed, we can predict loads exerted on offshore engineering structures. Each arrow represents a phenomenon which is simulated by an interaction-agent that computes its own contribution to the force balance on the system constituents.

to keep real-time intention. Thus we want to put into action our co-simulation architecture in order to abstract some ice behaviors and foresee the loadings evolution.

Marchenko model [24] provides a macroscopical description of the loading process on the structure based on observations. As the ice sheet moves forward, ice fragments pile up and form a ridge close to the platform. The part of the ridge which is above waterline is called the sail, and the part which is below waterline is called the keel, as Figure 6 illustrates. The interaction of the ice with the wall has a cyclic form that can be divided into two stages. In the first stage, the sail grows up fed by new ice fragments while the keel does not change. In the second stage, when the hydrostatic equilibrium between the sail and the keel is broken, the ice sheet fails and a part of the sail is transferred to the keel, and a sharp fall in the ice load on the wall is observed.

This model has the advantage to yield good results in load estimation for inclined structure while ensuring short computation time. But, as a semi-empirical one, it also dictates a number of strong assumptions which are not necessarily true for all situations. For example, it considers some parameters (sail and keel slopes, and porosities) as constants throughout the simulation. While it can be seen as true for most cases, this hypothesis becomes unrealistic for some particular ones. Hence, we would like to use Ice-MAS as a microscopical model to guide the Marchenko model during its simulation.

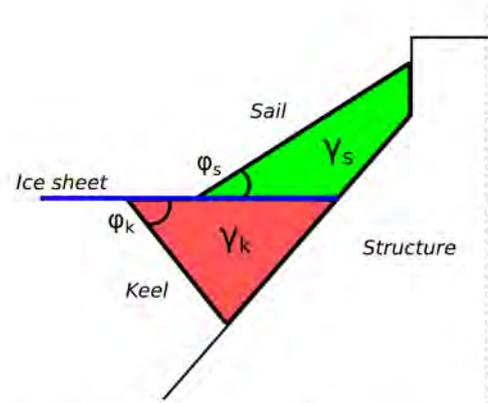


Fig. 6. Side view of ice-loading process. The advance of the ice sheet leads to the formation of a ridge in front of the structure. Based on observations during the course of experiments in an ice tank, we define the ridge formation process as follows. Ice fragments pile up above the waterline to form the sail. When the hydrostatic equilibrium is broken, the ice sheet fails and a part of the sail is transferred to the keel. In Marchenko model, it is assumed that slope of the sail  $\phi_s$  and the slope of the keel  $\phi_k$  remain unchanged as the ice floes moves onto the wall, as well as their respective porosities  $\gamma_s$  and  $\gamma_k$ .

## B. Results

We have to regularly recalibrate a set of four parameters ( $\phi_s, \phi_k, \gamma_s, \gamma_k$ ) of the Marchenko model, in order to obtain a more realistic macro-scale simulation. We apply therefore our implicit co-simulation strategy. Three simulations of the Marchenko model are run in parallel with three different sets of parameters. An Ice-MAS simulation is initialized according to the current Marchenko state. It is then run for a short duration on a fourth core. We use the vertical load on the structure, as both a result of Ice-MAS and Marchenko model, to choose the better macro-scale simulation. The operation is repeated for each cycle of the loading process, i.e. every time the hydrostatic equilibrium is broken. Figure 7 illustrates the solution construction, while Figure 8 shows a comparison between a full Ice-MAS simulation, the original Marchenko simulation with constant parameters and our adaptive Marchenko simulation.

## V. CONCLUSION

The architecture that we have proposed permits the fast computation of a macro-level simulation, in our case based upon a phenomenological model, with an extra accuracy provided by a micro level simulation. This is done by running concurrently many instances of the macro level simulator, running with distinct parameters set, and an instance of a micro level simulator. Since our phenomenological simulators are not parallelizable, or at least not easily parallelizable, they cannot, alone, take advantage of nowadays multi-core architecture. We cannot gain computation time thanks to extra cores, but at least, thanks to this architecture, we can therefore use extra cores to gain accuracy, without any additional computation delay. In the case of offshore structures design for Arctic conditions, our architecture allows us to build a fast prototyping simulator, using Marchenko model for the macro-level computation, and Ice-MAS as a micro level simulator. The obtained prototyping simulator gives very fast results, with an accuracy more acceptable than a simple usage of a plain Marchenko model.

## ACKNOWLEDGMENT

The authors express their appreciation to A. Dudal (Bureau Veritas) and B. Roberts (Technip) for technical expertise and support in modelling of ice behaviour, analysis of ice data and visualisation of ice interaction with offshore structures.

## REFERENCES

- [1] M. Wooldridge, "Intelligent agents : Theory and practice," *Knowledge Engineering Review*, 2001.
- [2] S. Kerdélo, "Méthodes informatiques pour l'expérimentation in virtuo de la cinétique biochimique. application à la coagulation du sang," Ph.D. dissertation, Université de Bretagne occidentale-Brest, Jan. 2006.
- [3] C. Le Gal, M. Olagnon, M. Parenthoen, P.-A. Beal, and J. Tisseau, "Comparison of sea state statistics between a phenomenological model and field data," in *OCEANS 2007 - Europe*. IEEE, 2007, pp. 1–6.
- [4] M. Combes, C. Grigné, L. Husson, C. P. Conrad, S. Le Yaouanq, M. Parenthoen, C. Tisseau, and J. Tisseau, "Multiagent simulation of evolutive plate tectonics applied to the thermal evolution of the Earth," *Geochemistry, Geophysics, Geosystems*, vol. 13, no. 5, p. Q05006, 2012, pL02926 PL02926.
- [5] P. Redou, G. Desmeulles, J. Abgrall, V. Rodin, and J. Tisseau, "Formal validation of asynchronous interaction-agents algorithms for reaction-diffusion problems," in *PADS'07, 21st International Workshop on Principles of Advanced and Distributed Simulation. In virtuo Experiments Based on the Multi-Interaction System*, vol. 329, 2007.
- [6] P. Redou, L. Gaubert, G. Desmeulles, P. A. Bal, C. Le Gal, and V. Rodin, "Absolute stability of chaotic asynchronous multi-interactions schemes for solving ODE," *Computer Modeling in Engineering and Sciences*, vol. 70, no. 1, p. 11, 2010.
- [7] G. Desmeulles, S. Bonneaud, P. Redou, V. Rodin, and J. Tisseau, "In virtuo experiments based on the multi-interaction system framework: the RISCOP meta-model," *Computer Modeling in Engineering and Sciences (CMES)*, vol. 47, no. 3, p. 299, 2009.
- [8] L. Crépin, "Couplage de modèles population et individu-centrés pour la simulation parallélisée des systèmes biologiques. application à la coagulation du sang," Ph.D. dissertation, Université de Bretagne occidentale-Brest, 2013.
- [9] S. Le Yaouanq, P. Redou, C. Le Gal, J. F. Abgrall, and J. Tisseau, "Multi-agent systems and heterogeneous scales interactions. application to pharmacokinetics of vitamin k antagonists," *Advances in Artificial Life, ECAL 2011*, 2011.
- [10] A. Fick, "Ueber diffusion," *Annalen der Physik*, vol. 170, no. 1, pp. 59–86, 1855.
- [11] S. Delaux, C. L. Stevens, and S. Popinet, "High-resolution computational fluid dynamics modelling of suspended shellfish structures," *Environmental Fluid Mechanics*, vol. 11, no. 4, pp. 405–425, 2010.
- [12] G. Pavliotis and A. Stuart, *Multiscale methods: averaging and homogenization*. Springer, 2008, vol. 53.
- [13] I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, C. Theodoropoulos *et al.*, "Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis," *Communications in Mathematical Sciences*, vol. 1, no. 4, pp. 715–762, 2003.
- [14] G. Samaey, I. G. Kevrekidis, and D. Roose, "Patch dynamics: Macroscopic simulation of multiscale systems," *PAMM*, vol. 7, no. 1, pp. 1 025 803–1 025 804, 2007.
- [15] P. Ming and J. Z. Yang, "Analysis of a one-dimensional nonlocal quasi-continuum method," *Multiscale Modeling & Simulation*, vol. 7, no. 4, pp. 1838–1875, 2009.
- [16] E. Weinan, B. Engquist, and Z. Huang, "Heterogeneous multiscale method: a general methodology for multiscale modeling," *Physical Review B*, vol. 67, no. 9, p. 092101, 2003.
- [17] A. Abdulle, E. Weinan, B. A. Engquist, and E. V. Eijnden, "The heterogeneous multiscale method," *Acta Numerica*, vol. 21, pp. 1–87, 2012.
- [18] R. Duboz, "Intégration de modèles hétérogènes pour la modélisation et la simulation de systèmes complexes," *Application à la modélisation multi-échelles en écologie marine*, 2004.
- [19] T. Ralston, "Plastic limit analysis of sheet ice loads on conical structures," in *Physics and Mechanics of Ice*. Springer, 1980, pp. 289–308.
- [20] Nevel, "Ice forces on cones from floes," in *IAHR-92*, vol. 3, 1992, pp. 1391 – 1404.
- [21] I. Konuk, A. Gurtner, and S. Yu, "Cohesive element framework for dynamic icestructure interaction problems. part II : implementation," in *Proceedings of OMAE 2009*, 2009.
- [22] R. Lubbad and S. Loset, "A numerical model for real-time simulation of ship-ice interaction," *Cold Regions Science and Technology*, vol. 65, no. 2, pp. 111–127, 2011.
- [23] C. Septseault, P.-A. Béal, S. Le Yaouanq, A. Dudal, and B. Roberts, "A New Ice SimulationTool Using a Multi-Model Program," *Artic Technology Conference 2014*, 2014.
- [24] A. Marchenko, "A method of calculating ice loads when ice piles up on a fixed wall," *Journal of Applied Mathematics and Mechanics*, vol. 70, no. 3, pp. 387 – 398, 2006. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0021892806000591>

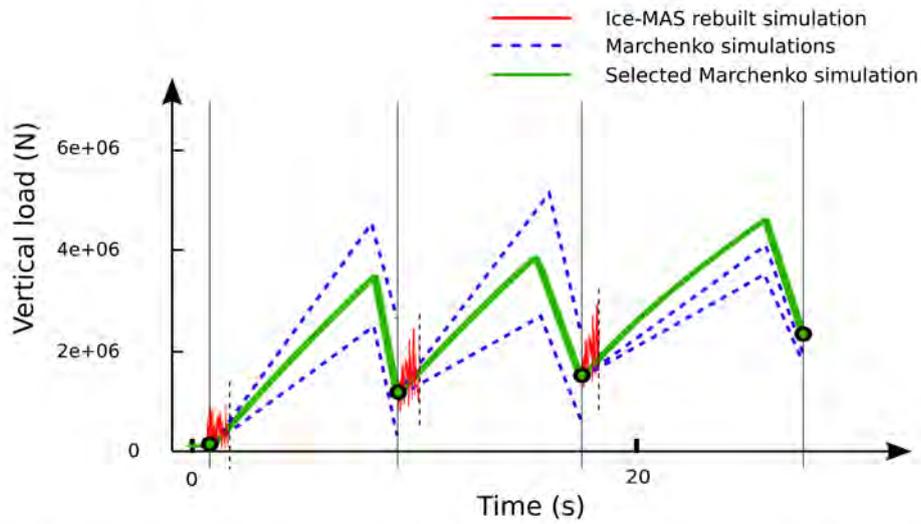


Fig. 7. Step-by-step construction of a solution thanks to co-simulation. For each loading cycle, three instances of Marchenko model are launched in parallel with different parameters sets. We can see that the three simulations thus lead to different results. In a fourth core, we initialize Ice-MAS according to the current macro-scale state and this micro-scale simulation is run for a shorter computation time. The best Marchenko simulation is selected by comparing the load slope from Ice-MAS.

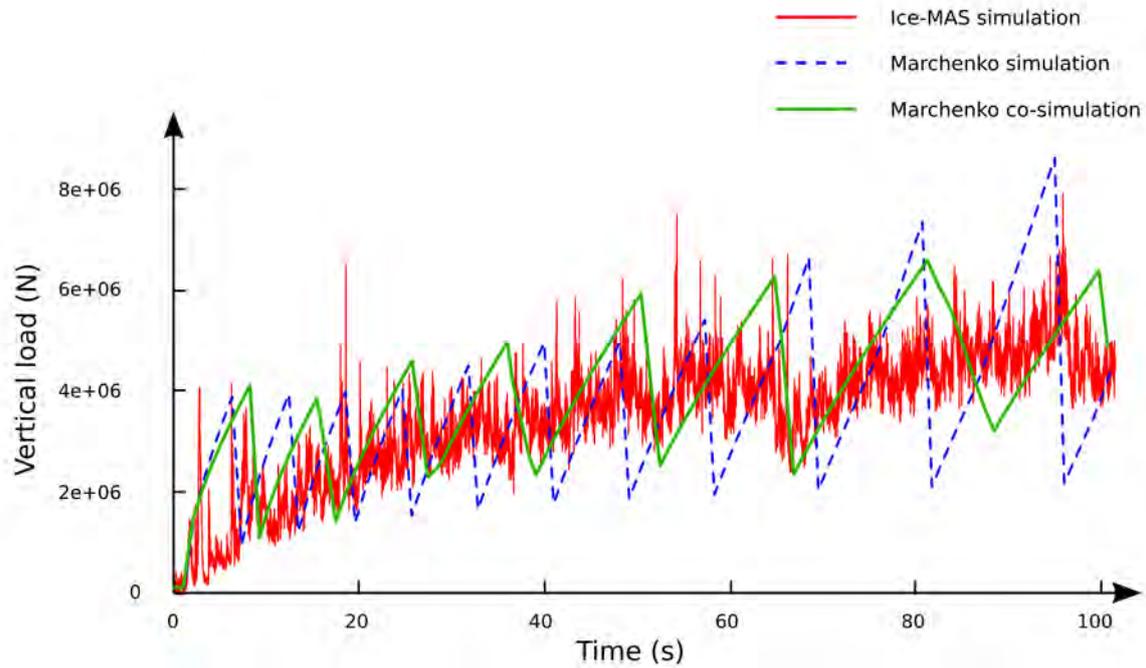


Fig. 8. Validation of co-simulation results. The auto-adaptive Marchenko solution shows better agreement than the standalone Marchenko model, which tends to get too large amplitude. This result follows an implicit computation of parameters which were considered previously as constants throughout the whole simulated time.

# Math modelling of the basic defensive activities

Jan Mazal<sup>1</sup>, Petr Stodola<sup>2</sup>, Libor Kutěj<sup>3</sup>, Milan Podhorec<sup>4</sup>, Dana Křišťálová<sup>5</sup>

<sup>1</sup> Faculty of Economics and Management, University of Defence, Kounicova 65, 662 10 Brno, Czech Republic.

E-mail: jan.mazal@unob.cz

<sup>2</sup> Faculty of Economics and Management, University of Defence, Kounicova 65, 662 10 Brno, Czech Republic.

E-mail: petr.stodola@unob.cz

<sup>3</sup> Ministry of Defence, Tychonova 1, 160 01 Prague, Czech Republic. E-mail: libor.kutej@centrum.cz

<sup>4</sup> Faculty of Economics and Management, University of Defence, Kounicova 65, 662 10 Brno, Czech Republic.

E-mail: milan.podhorec@unob.cz

<sup>5</sup> Faculty of Economics and Management, University of Defence, Kounicova 65, 662 10 Brno, Czech Republic.

E-mail: dana.kristalova@unob.cz

**Abstract** – Modelling support of tactical tasks is not exceptional by these days, but it is not the part of the direct decision support of the security managers or commanders in the operations yet. In context of the latest trends of technology development and requirements on C4ISR (military information systems – command, control, communication, computer, intelligence, surveillance, reconnaissance), the future of operational and technological development of 21-st century battlefield is moving to the real time modeling and simulation approaches in military decision making process support. This problematics contain math modelling of operational tasks, frequently multi-criteria decision problems, supported by the latest information technology. This article is dedicated to the topic of a future battlefield, focused on math solution and optimization of a force maneuver and deployment in various operational activities. The solution of a problem count with the multi-criteria optimization on a huge operational data set, problematic of the tactical requirements or criteria quantification, math modelling of operational activities and optimization function solution, search for the extremes and the stability analyses.

**Keywords** – C4ISR, Decision Support Systems, Optimal Deployment, Optimal Maneuver, MDMP.

## I. INTRODUCTION

One of the most important capability in the military or security applications is the fast and rational decision making process. Decision-making activities by these days could be supported by advanced theory and technology, were the application attempts of mathematical modeling in the military art or security applications have been known for centuries.

Current decision making process in military or security environment is similar to its civilian corresponding counterpart, but with different inputs, outcomes and consequences. The commanders or security managers are searching for optimal multi-criteria solution, mostly balanced with some contradictory requirements and respecting relevant factors like: time (quickness of decision making process), the issue of accessible resources, unfamiliar environment (territory, opponent, other inhabitants, technology) and mainly the acceptable risk level of the friendly forces involvement.

The increasing dynamics of the future conflicts will impose a strong requirement on a decision making process of the commanders or security managers to make the decisions quickly and rationally with highest pragmatic impact on a concrete operational situation. By these days, the most of the key decisions in (military/security) operations are established on intuition and experience (empirical-intuitive decision-making process). Because of that fact, we could test the implementation of a wide set of optimization methods based on mathematical modeling and simulation approaches.

## II. THE APPROACH

Math modeling and simulation is widely applied in many areas of industry and trade sector at that time, implementing the methods from operations research, especially a linear or stochastic programming, used in business planning and strategy modelling (what-if analysis, business scenario analysis). Modelling decision support of military or security applications and procedures is not exceptional today, but it still falls within the range of the direct decision support in real operations.

Historically, the first “advanced” approaches to mathematically modelate the combat activities were carried out in the 1960s, mostly dedicated to the operational tasks related to the Cold War. The math models issued by these days were based on a very general assumptions and tried to construct the rationality of the certain entity behavior in the very approximate terms.

It should be mentioned that the original math models (mostly based on the sets of several differential equations) were developed in context of the available technology (i.e. low computational power and lack of complex operational databases) and took into account insufficient amount of information (several coefficient from the very large area of the battlefield), because of that, it prevented them from incorporating a sufficient level of detail, necessary for a practically acceptable results.

From philosophical point of view, it is possible to split the concept of modelling (computer) support of decision making process in operational environment into two lines, namely:

- Subjective - empirical and intuitive

Contemporary decision-making process of the many security managers or commander is still executed in terms of experience and intuition and probably it will keep this character in the near future.

- Objective – mathematical and algorithmic

Mathematical support within algorithmic (computer) approach is still a relatively new approach which, even though some initial attempts of its "start-up" done in the past, is still on the beginning and probably it takes some time to accommodate that "philosophical upgrade" in the decision-making activities of the security managers and commanders mainly on the tactical level. For effective "operational" decision making, it is beneficial to keep the coexistence of both approaches in the balanced interaction and complementarity in such a ratio that comply with the type of the specific decision-making problem.

As it was mentioned before, the initial math models have suffered from a serious deficiencies related to a sufficient amount of operational information, what the new operational decision modelling concept should improve. Major upgrade of a new approach in context of previous solutions brings new aspect, which consist in:

- Comprehensive data-structured concept of the operational environment.
- Detailed real-time virtualization of the operational area.
- Extensive extrapolation of operational attributes (status) in wide range of situations.
- Advanced operational and tactical analyses, integrated into math models and final solutions, respecting the multi-criteria requirements.
- Sharing operational information in real time – the fast dissemination of the current (status, attributes and so on) information from the operational environment is undoubted vital for effective decision making. This fact was already proved in the last decade of the military conflicts and it creates for example the fundamentals of the modern C4ISR (Command, Control, Communication, Computer, Intelligence, Surveillance, Reconnaissance) and FSS (Future Soldier Systems) systems.
- Expert systems – include decision trees, models based on fuzzy logic, etc. This systems are common in the industry and business sector but in the security or military area it is still not too frequent yet. In operational modeling it hides a great potential.

Leading position in the area of advanced and automated modelling support of operational decisions still keeps the US military. US introduced the revolutionary operational and tactical approach called the Deep Green concept [13]. Deep Green concept was inspired by a success and philosophy of a Deep Blue supercomputer (1997) and it is focused on a real time solution of advanced operational and tactical tasks dedicated to the future military operations on the battlefield of 21st century. Deep Green concept is a project issued by the

DARPA (Defence Advanced Research Project Agency) in 2008.

## OPTIMAL POSITION IN DEFENCE

### A. Motivation

Generally, the search for the optimal position in the defence activities is a very complex and demanding problem, if we want to take in account all aspects of the real operating environment. Because of that fact, certain initial approximation and simplification is necessary. For the demonstration of the basic approach to that issue, there are following assumptions and conditions:

There is one friendly tactical entity in a source area and one destination area where enemy tactical entity could appear.

We expect, that advancing enemy entity will attack friendly entity in the source area with ability to take some damage the enemy entity will be able to advance in destination area, because of that, the destination area is splitted into two parts (primary and secondary).

Task - the friendly tactical entity is required to find two suitable locations (each for the first and second part of the destination area) in order to shoot (destroy or disable) the enemy entity.

Task conditions – friendly tactical entity should minimize its own exposure to the enemy entity between the movement of the two positions and each defensive position must fulfill the best tactical condition for the shooting position (in a defence) - defined by the balanced combination of the distance to the target, ability to hit the target and ability to conceal and take cover for the friendly entity.

### B. Analysis

In order to formalize the tactical situation, we assume a mathematical structure  $M$  which represents the operational environment (the set of all possible locations). Generally,  $M \subset \mathbb{R}^3$  (three dimensional space). When restricted to ground operations, the terrain can be viewed as a mapping  $g : M \rightarrow \mathbb{R}^3$ , where  $M \subset \mathbb{R}^2$ . Alternatively, a graph structure  $G = (M, E)$  can be used to model the surface (terrain) maneuver, where  $M$  is the set of nodes.

Further, we assume that the enemy entity will appear in the primary area, attacking the friendly entity and advancing to the secondary area (continuing with the attack). At that case, we look for the two suitable positions for the friendly entity (for the each part of destination area), fulfilling the condition of the most safe maneuver between them. Desired optimization aim is expressed by the following formula:

$$\max (Fsp(Dx_1, A_1, CvA_1) + Fsp(Dx_2, A_2, CvA_2) + M(x_1, x_2)); \quad (1)$$

Where:

$Fsp()$  final pragmatism of fire (shooting pragmatism, linked to a particular position);

$M()$  final pragmatism of maneuver (between  $x_1$  and  $x_2$ );

- $x_1, x_2$  positions in area 1 and 2;
- $Dx_1$  distance to the target in area 1 from position  $x_1$ ;
- $Dx_2$  distance to the target in area 2 from position  $x_2$ ;
- $A_1, A_2$  the difference of the excess of friendly and enemy entity in the area 1 and 2;
- $CvA_1, CvA_2$  distance to the closest cover in area 1 and 2;

At the first, the overall goal in choosing the best position  $x$  in source area is to maximize the chance of hitting the target at any position in each destination area, while minimizing self exposure to the target. There are many criteria on the shooting position which relate to this goal, but initially for that example were chosen follows:

- Position accessibility,
- Visibility of the target,
- Position with respect to the target (e.g. distance, elevation),
- Camouflage properties of the location (e.g. vegetation, prevailing color, etc.)
- Maneuver to the closest cover.

The notion of pragmatic aspect just in that model, refers to the position's overall suitability under the above-described conditions. The number of multi-criteria conditions imposed on that task could be many, but approximation is necessary in that "initial" modeling, because each input increase the dimension of the partial model by one and most of these complex models needs further experimental testing and evaluation.

The shooting model (3) was inspired by the next formula (2) [15], the graph is shown on Figure 1:

$$f(x, y) = \left(\frac{2}{3} \operatorname{atan}\left(\frac{0,5y-30}{x+1}\right) + 1\right) \left(\frac{810}{\frac{x}{60}+1}\right) \left(0,9 - \frac{1}{\frac{(x+15)}{100}+1}\right); \quad (2)$$

Where:

- $f(x,y)$  final pragmatism of fire;
- $x$  distance to the target,  $x \in R \cap (0,500)$ ;
- $y$  the difference of the excess of friendly and enemy entity,  $y \in R \cap (-80, 80)$ ;

Detailed description of all input criteria and its achievement (as an important part of the integration model to quantify the input characteristics), would significantly exceeded the framework of the article. Nevertheless, in general overview, it is set of the models applying a wide spectra of algorithms and multi-dimensional functions.

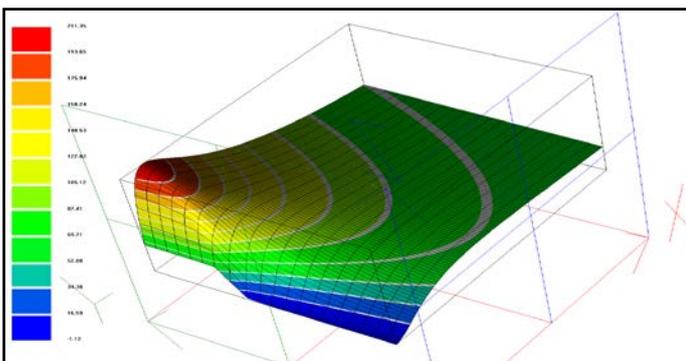


Figure 1: Graph of a shooting model (2)

Main aim of the operational modeling at that case is to incorporate the influence of external laws, conditions and characteristics to the numeric set of pragmatically coefficients (define the level of pragmatism of desired activity under the considered conditions), which are applied to the final model development. Therefore, as an example may serve the following fire pragmatism formula<sup>1</sup> defining the pragmatism coefficient of the entity position in the context of the contact with enemy entity. General function  $F_{sp}(n1,n2,...,nm)$  was limited in this case by the inputs  $m = 3$ .

$$f_{sp}(x, y, w) = \frac{0,51(5-w/10)}{2} \frac{(155 \operatorname{atan}\left(\frac{y}{x}\right)+200)}{\left(\frac{x-50 \operatorname{atan}\left(\frac{y}{x}\right)+40(5-w/10)}{90}-3\right)^2} + \frac{3(5-w/10) \operatorname{atan}\left(\frac{y}{x}\right)}{2} + 1 \quad (3)$$

- $f_{sp}(x,y,w)$  final pragmatism of fire;
- $x$  distance to the target (10 – 500);
- $y$  the difference of the altitude of the entities (- 150, 150);
- $w$  length of the path to the closest cover or vegetation (0,50);

Figure 2 shows the intuitively derived mathematical model of the fire pragmatism with 3 (distance to the target, the difference of the altitude of the entities, length of the path to the closest cover) selected parameters. The  $x$  axis represents the distance to the target in the model range of 10-500 meters, the  $y$  axis represents elevation difference of the entities. Because the model (function) takes three inputs (variables), its dimension is equal to four (inputs increased by one). Its overall representation in the 3D view is problematic, so on Figure 2, there is a presentation of the 3D cuts of the 4D model by a particular parameter (input), in that case, the cuts are made according to the parameter  $w$  (length of the path to the nearest cover or vegetation), see Figure 2.

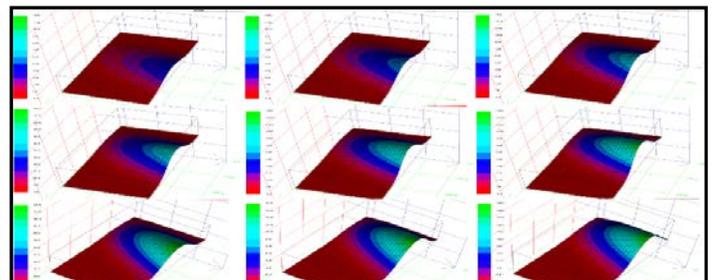


Figure 2: 3D cut of 4D model according to the expression (3), parameter  $w$  is in range of 45-0m in steps of 5m.

The solution takes in account the position of the attack versus the position of the target (enemy entity) and maneuver optimality to the next attacking position (derived from the position matrix of the strike pragmatism). The construction of  $M$  from the individual criteria can be carried out in the way, where the coefficients for the maneuver pragmatism are calculated by modified Floyd-Warshall algorithm for all

<sup>1</sup> Model of the fire pragmatism is essentially an integration of multiobjective inputs and variation of conditions. Generally, the model can integrate any number of variables, each input/variable increases the dimension of the model and its complexity. Models of higher dimensions can be visualize only in their cuts (slices).

positions of possible turn in the model of operational environment. The construction of M from the individual criteria can be carried out in a following way:

$$M(x, y) = Cx - \text{MinPathCost}(x, y) \quad (4)$$

The parameters are:

- $x$  location of friendly tactical entity in the source area 1
- $y$  location of friendly tactical entity in the source area 2
- $Cx$  maximal pragmatically coefficient

Theoretical approach to the tactical maneuver modeling is illustrated on Figure 3 and is split into three phases implementing complex operational area database and models, geographical and tactical analyses, enemy and friendly tactical entities ability estimation and optimized dynamic programming algorithms for fast (computer) processing on a large datasets (terrain models are about >100MB, divided into slices of attribute matrixes of 2048 x 2048).

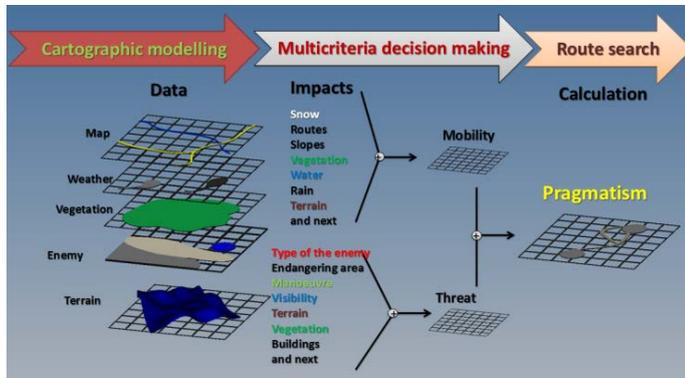


Figure 3: Theoretical approach to the optimal maneuver solution

Layout of pragmatically coefficients in one particular path are shown on the following picture:

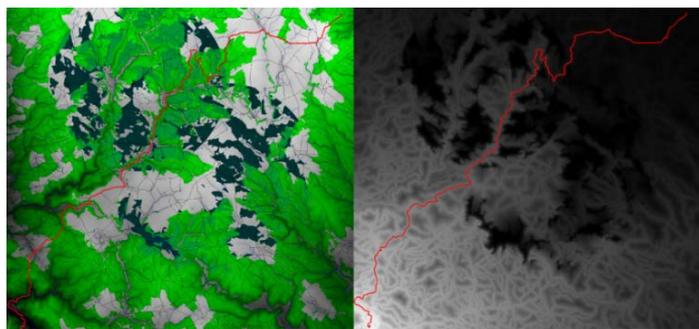


Figure 4: Original mapp and Graph pragmatism visualization

Resulting matrix of the strike pragmatism is integrated with a matrix of a maneuver pragmatism to the final “optimality” matrix containing the pragmatism coefficient for the defence positions in mentioned context. Optimization process (implementing expression (1)) is iteratively executed and is carried out for all possible combination of the friendly entity position in each source area. After all iterative steps, the computer searches in the result database for the highest pragmatically coefficient of all particular solutions.

After that step, the perspective solutions is further analyzed, especially for its stability. If the solution does not comply with

the conditions of stability (isolated peak of pragmatism), so another potential solution is selected from the database and sent to the same analysis. The first solution that meets the criteria of stability and optimality is presented to the user as a possible configuration of the friendly tactical entity positions, optimal maneuver and location of enemy entity.

As an example may serve the solution illustrated on Figure 5, where the orange squares represents the extrapolated areas of the advancing opponent and the blue circles indicates the position of friendly elements. The optimal maneuver between the optimal positions for each area is marked in red.

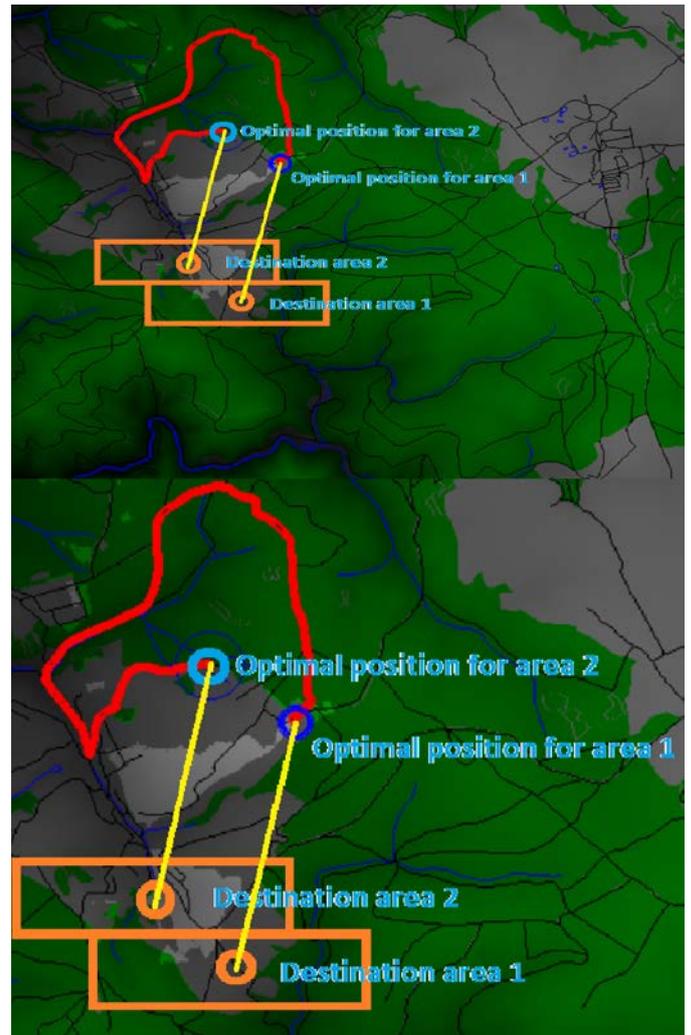


Figure 5: The defence position optimization

### III. CONCLUSION

If we look at the fundamental purpose of the security or military organizations and its orientation to the combat activities, it is easy to derive, that the decision making process of the security managers or commanders usually follow the pragmatic concept of optimization of the specified tactical activity (or sequence of activities), issuing in task competition, for example in the shortest time, minimum effort, minimum losses, minimal resources, with maximal safety, etc. Like it was demonstrated in presented example of optimal defensive “behavior” under the certain conditions.

Modelling and simulation of operational tasks, even it is not

apparent at the first look, it follows the pragmatic concept of operation and enable implementation of math-algorithmic approach and further automation. The solutions of operational problems are not usually simple and the results are sensitive on the input data precision and set of the multi-criteria requirements. Also, the final results is usually necessary to further analyze in terms of their stability.

Solution of the particular operational task is based on individual approaches and could not be generalized. The overall concept should be perceived as a complex problematic, rather than standalone problem. At that time, there exists no universal solution to address more different operational tasks. The solution of individual operational problem usually address the multi-criteria integration of operational analysis and models linked to the proper quantification and criteria setting.

Despite the fact that the current modeling of the operational activities is from the philosophical point of view relatively highly theoretical matter and it is still on the beginning, it is intuitively obvious that the future potential of mentioned models and its practical application can be very high. This approach is upgrading a proved, but static concept of real-time data dissemination to a new dimension and could serve as a powerful tool in the planning and operation management phase.

#### REFERENCES

- [1] Mokhtar S. Bazaraa and John J Jarvis. *Linear programming and network flows* / Mokhtar S. Bazaraa, John J. Jarvis. Wiley, New York :, 1977.
- [2] A. Bondy and U.S.R. Murty. *Graph Theory*. Graduate Texts in Mathematics. Springer, 2008.
- [3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Berichte uberverteilte messsysteme. Cambridge University Press, 2004.
- [4] George Dantzig. *Linear Programming and Extensions*. Landmarks in Physics and Mathematics. Princeton University Press, 1998.
- [5] J.D. Foley. *Computer Graphics: Principles and Practice*, Second Edition in C. The Systems Programming Series. Addison-Wesley Pub, 1996.
- [6] G. Ghiani, G. Laporte, and R. Musmanno. *Introduction to Logistics Systems Planning and Control*. John Wiley and Sons, Ltd, 2004.
- [7] F. Glover and M. Laguna. *Tabu Search*. Number sv. 1 in *Tabu Search*. Kluwer Academic Publishers, 1998.
- [8] P. Kall and S. W. Wallace. *Stochastic Programming*. John Wiley and Sons, Chichester, second edition, 1994.
- [9] G.J. Klir and B. Yuan. *Fuzzy sets and fuzzy logic: theory and applications*. Prentice Hall PTR, 1995.
- [10] M. Kress. *Operational Logistics: The Art and Science of Sustaining Military Operations*. Springer, 2002.
- [11] C.R. Rao and H. Toutenburg. *Linear Models: Least Squares and Alternatives*. Springer Series in Statistics. Springer, 1999.
- [12] Mikulas Rybar. *Modelovanie a simulacia vo vojenstve*. Ministerstvo obrany Slovenskej republiky, Bratislava, 2000.
- [13] J.R. Surdu and K. Kittka. *The deep green concept*. In *Spring Simulation Multiconference 2008 (SpringSim'08), Military Modelling and Simulation Symposium (MMS)*, 2008.
- [14] A. Washburn and M. Kress. *Combat Modeling*. International Series in Operations Research & Management Science. Springer, 2009.
- [15] I. Mokra. *Modelovy prstup k rozhodovcm aktivitm velitelu jednotek v bojvch operacch*. Disertan prce. Brno: Univerzita obrany v Brn, Fakulta ekonomiky a managementu, 2012. 120 s.
- [16] Blha, M.; Brabcova, K. *Decision-Making by Effective C2I system*. In: *7th International Conference on Information Warfare & Security*. Seattle: Academic Publishing Limited, 2012, p. 44-50. ISSN 2048-9870. ISBN 978-1-908272-29-4.

# Polarization-Insensitive Perfect FSS Metamaterial Absorber in THz Frequency Range

C. Sabah, F. Dincer, M. Karaaslan, E. Unal and O. Akgol

**Abstract**— We numerically presented and analyzed a new perfect frequency selective surface (FSS) metamaterial absorber (MA) based on resonator with dielectric configuration for terahertz frequency ranges. Proposed FSS MA has features of simple configuration and easy fabrication. Also, it introduces flexibility to adjust its FSS metamaterial (MTM) properties and easily re-scale the model for various other frequencies. Moreover, numerical simulations verify that the FSS MAs could achieve very high absorption at wide different all polarization angles. The proposed FSS MAs and its variations enable myriad potential application areas in defend systems, communication, stealth technologies, and so on.

**Keywords**— absorber; metamaterial; terahertz.

## I. INTRODUCTION

MTMs are artificially created electromagnetic (EM) materials have gained great attention of science community. Since, MTMs show specific EM features not ordinarily encountered in nature such as negative refractive index [20, 12, 31, 5, 6]. Also, MTMs are manmade and have many potential application areas for example cloaking [3], absorber [8], super lens [10], sensing [26], antenna [22], and so on [9, 21, 25, 29, 1, 19, 14, 4].

Nowadays, the concept of MA studies has gained attention by the scientists who study on MTMs. There are many MA studies in literature. These studies are commonly realized on microwave regime. However, researchers studied on also ranges of THz and infrared frequency in last few years. Some of these are broadband terahertz absorber [13], multi-band THz MA [11], polarization-independent plasmonic absorber [7], broadband MA [24].

We considered and analyzed on the MA studies in literature. Unlike the others, we presented perfect FSS MA that operates in terahertz frequency ranges and has easy fabrication

techniques. Also, we are investigated with respect to dependency on polarization angles of the suggested model. Moreover, the proposed FSS MA model has comfortable configuration and can easily be re-scaled for other frequencies. The proposed FSS MA and its variations enable numberless potential applications in medical technologies, sensors, wireless communication, and so on.

## II. THEORETICAL APPROACH

The frequency response of absorption is defined as  $A(\omega)=1-R(\omega)-T(\omega)$ , where  $A(\omega)$ ,  $R(\omega)$  and  $T(\omega)$  are the absorption, reflectance and transmittance, respectively.  $A(\omega)$  comes from minimizing either reflectivity  $R(\omega)=|S_{11}|^2$  and transmission  $T(\omega)=|S_{21}|^2$  at an specified frequency range.

Reflectivity can be reduced (near-zero) when the effective permittivity  $\tilde{\epsilon}(\omega)$  and permeability  $\tilde{\mu}(\omega)$  have the same value. It is possible to absorb both the incident electric and magnetic field tremendously by accurately tuning  $\tilde{\epsilon}(\omega)$  and  $\tilde{\mu}(\omega)$ . They can be manipulated to create high absorption. Absorbers minimize the reflection and transmission coefficients of incident waves at a certain frequency range due to the impedance matching [8]. In the resonance condition, the effective impedance  $(Z(\omega)=\sqrt{\tilde{\mu}(\omega)/\tilde{\epsilon}(\omega)}=z_1+iz_2)$  have to match with the free space impedance  $Z(\omega)=Z_0(\omega)$  and therefore, the reflection is minimized [16, 5, 23, 24, 18, 9].

## III. NUMERICAL STUDY, RESULTS, AND DISCUSSION

Proposed FSS MA design is based on square and rectangular-shaped inclusions. The models consist of a resonator, metallic layer and dielectric substrate. Resonator and metallic layer are modelled as silver sheet with electrical conductivity of  $6.3 \times 10^7$  S/m and thickness of 1  $\mu\text{m}$ . Silver is soft, white, lustrous transition metal and also possesses the highest electrical conductivity inside of metals. Also, it has extremely low resistivity. Resonator and metallic plate are separated by the Quartz (Fused)-dielectric substrate and placed parallel to each other. The thickness, loss tangent, relative permittivity and permeability of the Quartz (Fused) are 100  $\mu\text{m}$ , 0.0004, 3.75 and 1, respectively. Fig. 1 shows the structure designs with their dimensions.

M.K. acknowledges the support of TUBITAK under the Project Number of 113E290 and partial support of the Turkish Academy of Sciences

C. Sabah is with the Department of Electrical and Electronics Engineering, Middle East Technical University - Northern Cyprus Campus, Kalkanli, Guzelyurt, TRNC / Mersin 10, Turkey (e-mail:sabah@metu.edu.tr).

F. Dincer is with the Department of Computer Engineering, Mustafa Kemal University, Iskenderun, Hatay, 31200, Turkey

M. Karaaslan, E. Unal and O. Akgol are with the Department of Electrical and Electronics Engineering, Mustafa Kemal University, Iskenderun, Hatay, 31200, Turkey.

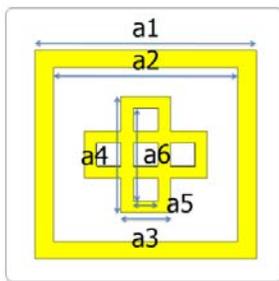


Fig. 1 Dimensions of the suggested FSS MA for terahertz frequency range- $a_1= 360$  um,  $a_2=300$  um,  $a_3=80$  um,  $a_4=200$  um,  $a_5=40$  um,  $a_6=160$  um

We performed a commercial full-wave EM solver based on the finite integration technique for numerical studies of the periodic structure. So, we used the periodic boundary conditions with floquet port. Then, we numerically analyzed and compared results to obtain characteristics of the others FSS MA. The FSS MA shows perfect single band around 0.99 THz in the reflection spectrum thus perfect single maxima in the absorption as shown in Fig. 2(a). The resonance is about 99.98 % in the simulation. As seen, the amplitude of the reflection is 0.01 at the resonance frequency in this case.

In the next exploration, the effects of the polarization angle on the performance of the FSS MA are observed. Fig. 2(b) shows the frequency response of the absorption value for the stated process. To notice the shifts of the resonance frequency with respect to the polarization angle, wider frequency range is taken into consider as shown in Fig. 2(b). It can be seen that the proposed FSS MA provides very well absorption for  $0^\circ$ ,  $120^\circ$ ,  $150^\circ$  and  $90^\circ$  with the absorptions of 88.11 %, 98.07 %, 87.22 % and 99.98 %, respectively. The lowest absorption value is occurred around 1.05 THz as 46.95 % for  $60^\circ$  and the highest absorption is occurred around 0.99 THz as 99.98 % for  $90^\circ$ . Although the suggested MA does not provide good polarization angle independency, it has good resonance with small shifts for all incident angles except for  $60^\circ$ .

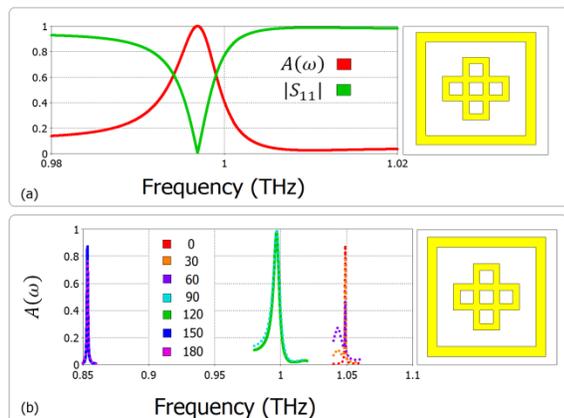


Fig. 2 Proposed THz FSS MA, a) Simulated reflection and absorption for the proposed terahertz absorber, (b) Simulated results of the absorbing performance under different polarization angles for the case of normal incidence.

## IV. CONCLUSION

In conclusion, THz FSS MA is presented. This proposed model shows perfect absorption and also it can be tuned easily re-scale the structure for other frequencies. According to the results, the investigated FSS MA provides perfect absorption at resonance frequencies independent from the polarization angles. At some cases of the polarization angles, the absorption value can be enhanced. Moreover, it can be a good candidate for the applications of stealth, sensor, and so on.

## REFERENCES

- [1] O. Akgol, D. Erricolo, and P. L. E. Uslenghi, "Electromagnetic radiation and scattering for a gap in a corner backed by a cavity filled with DNG metamaterial", *Radio Sci.*, vol. 46(4), 2011.
- [2] A. Alù, and N. Engheta, "Anomalies of subdiffractive guided wave propagation along metamaterial nanocomponents", *Radio Sci.*, vol. 42(6), 2007.
- [3] A. Alù, and N. Engheta, "Robustness in design and background variations in metamaterial/plasmonic cloaking", *Radio Sci.*, vol. 43(4), 2008.
- [4] S. Arslanagic, R. W. Ziolkowski, and O. Breinbjerg, "Analytical and numerical investigation of the radiation from concentric metamaterial spheres excited by an electric Hertzian dipole", *Radio Sci.*, vol. 42(6), 2007.
- [5] H. Cheng, S. Chen, H. Yang, J. Li, X. An, C. Gu, and J. Tian, "A polarization insensitive and wide-angle dual-band nearly perfect absorber in the infrared regime", *J. Opt.* vol. 14, pp. 085102, 2012..
- [6] Y. Cheng, and H. Yang, "Design Simulation and Measurement of Metamaterial Absorber", *Microw. Opt. Techn. Let.*, vol. 52, pp. 877-880, 2010.
- [7] S. Dai, D. Zhao, Q. Li, and M. Qiu, "Double-sided polarization-independent plasmonic absorber at near-infrared region", *Opt. Express*, vol. 21, pp. 13125-13133, 2013.
- [8] F. Dincer, M. Karaaslan, E. Unal, and C. Sabah, "Dual-Band Polarization Independent Metamaterial Absorber Based On Omega Resonator and Octa-Starstrip Configuration", *Prog. Electromagn.Res.* vol. 141, pp. 219-231, 2013a.
- [9] F. Dincer, C. Sabah, M. Karaaslan, E. Unal, M. Bakir and U. Erdiven, "Asymmetric Transmission of Linearly Polarized Waves and Dynamically Wave Rotation Using Chiral Metamaterials", *Prog. Electromagn. Res.*, vol. 140, pp. 227-239, 2013b.
- [10] N. Fang, H. Lee, C. Sun, and X. Zhang, "Sub-Diffraction-Limited Optical Imaging with a Silver Superlens", *Science*, vol. 308, pp. 534-537, 2005.
- [11] F. Hu, L. Wang, B. Quan, X. Xu, Z. Li, Z. Wu, and X. Pan, "Design of a polarization insensitive multiband terahertz metamaterial absorber", *J. Phys. D: Appl. Phys.*, vol. 46, pp. 195103, 2013.
- [12] R. B. Hwang, "Correlation between a negative group velocity and a slanted stop band in two-dimensionally periodic structures", *Radio Sci.*, vol. 41(1), 2006.
- [13] D. S. Kim, D. H. Kim, S. Hwang, and J. H. Jang, "Broadband terahertz absorber realized by self assembled multilayer glass spheres", *Opt. Express*, vol. 20, pp. 13566-72, 2012.
- [14] M. R. Khodja, and E. A. Marengo, "Comparative study of radiation enhancement due to metamaterials", *Radio Sci.*, vol. 43(6), 2008.
- [15] T. M. Kollatou, A. I. Dimitriadis, S. D. Assimonis, N. V. Kantartzis, and C. S. Antonopoulos, "A Family Of Ultra-Thin, Polarization-Insensitive, Multi-Band, Highly Absorbing Metamaterial Structures", *Prog. Electromagn. Res.*, vol. 136, pp. 579-594, 2013.
- [16] N. I. Landy, S. Sajuyigbe, J. J. Mock, D. R. Smith, and W. J. Padilla, "A Perfect Metamaterial Absorber", *Phys. Rev. Lett.*, vol. 100, pp. 207402-4, 2008.
- [17] L. Lu, S. Qu, H. Ma, F. Yu, S. Xia, Z. Xu, and P. Bai, "A Polarization-Independent Wide Angle Dual Directional Absorption Metamaterial Absorber", *Prog. Electromagn.Res.*, vol. 27, pp. 191-201, 2012.

- [18] J. Lee, and S. Lim, "Bandwidth-enhanced and polarisation-insensitive metamaterial absorber using double resonance", *Electron. Lett.*, vol. 47, pp.8-9, 2011.
- [19] N. G. Lehtinen, "A waveguide model of the return stroke channel with a "metamaterial" corona", *Radio Sci.*, vol. 47(1), 2012.
- [20] C. Sabah, "Left-handed Chiral Metamaterials", *Cent. Eur. J. Phys.*, vol. 6, pp. 872-878, 2008.
- [21] C. Sabah, "Multiband Metamaterials Based On Multiple Concentric Open-Ring Resonators Topology", *IEEE J. Sel. Top. Quant.*, vol. 19, pp. 8500808, 2013.
- [22] L. M. Si, W. Zhu, and H. J. Sun, "A Compact, Planar, and CPW-Fed Metamaterial-Inspired Dual Band Antenna", *IEEE Antenn. Wirel. Pr.*, vol. 12, pp. 305-308, 2013.
- [23] J. Sun, L. Liu, G. Dong, and J. Zhou, "An extremely broad band metamaterial absorber based on destructive interference", *Opt. Express*, vol. 19, pp. 21155-62, 2011.
- [24] L. Sun, H. Cheng, Y. Zhou, and J. Wang, "Broadband metamaterial absorber based on coupling resistive frequency selective surface", *Opt. Express*, vol. 20, pp. 4675-80, 2012.
- [25] I. V. Shadrivov, A. A. Zharov, N. A. Zharova, and Y. S. Kivshar, "Nonlinear left-handed metamaterials", *Radio Sci.*, vol. 40(3), 2005.
- [26] H. Tao, E. A. Kadlec, A. C. Strikwerda, K. Fan, W. J. Padilla, R. D. Averitt, E. A. Shaner, and X. Zhang, "Microwave and Terahertz wave sensing with metamaterials", *Opt. Express*, vol. 19, pp. 21620-6, 2011.
- [27] H. Tao, N. I. Landy, C. M. Bingham, X. Zhang, R. D. Averitt, and W. J. Padilla, "A metamaterial absorber for the terahertz regime: Design, fabrication and characterization", *Opt. Express*, vol. 16, pp. 7181-7188, 2008.
- [28] H. Tao, C. M. Bingham, D. Pilon, K. Fan, A. C. Strikwerda, D. Shrekenhamer, W. J. Padilla, X. Zhang, and R. D. Averitt, "A dual band terahertz metamaterial absorber", *J. Phys. D: Appl. Phys.*, vol. 43, pp. 225102, 2010.
- [29] H. Odabasi, and F. L. Teixeira, "Analysis of canonical low-profile radiators on isoimpedance metamaterial substrates", *Radio Sci.*, vol. 47(1), 2012.
- [30] W. Qin, J. Wu, M. Yu, and S. Pan, "Dual-band terahertz metamaterial absorbers using two types of conventional frequency selective surface elements", *Terahertz Science and Technology*, vol. 5, pp. 169-174, 2012.
- [31] R. W. Ziolkowski, and C. Y. Cheng, "Lumped element models of double negative metamaterial-based transmission lines", *Radio Sci.*, vol. 39(2), 2004.

# The method of Probabilistic Nodes Combination in simulation and modeling

Dariusz J. Jakóbczak

**Abstract**—Proposed method, called Probabilistic Nodes Combination (PNC), is the method of 2D curve modeling and handwriting identification by using the set of key points. Nodes are treated as characteristic points of signature or handwriting for modeling and writer recognition. Identification of handwritten letters or symbols need modeling and the model of each individual symbol or character is built by a choice of probability distribution function and nodes combination. PNC modeling via nodes combination and parameter  $\gamma$  as probability distribution function enables curve parameterization and interpolation for each specific letter or symbol. Two-dimensional curve is modeled and interpolated via nodes combination and different functions as continuous probability distribution functions: polynomial, sine, cosine, tangent, cotangent, logarithm, exponent, arc sin, arc cos, arc tan, arc cot or power function.

**Keywords**— handwriting identification, shape modeling, curve interpolation, PNC method, nodes combination, probabilistic modeling.

## I. INTRODUCTION

Handwriting identification and writer verification are still the open questions in artificial intelligence and computer vision. Handwriting based author recognition offers a huge number of significant implementations which make it an important research area in pattern recognition [1]. There are so many possibilities and applications of the recognition algorithms that implemented methods have to be concerned on a single problem. Handwriting and signature identification represents such a significant problem. In the case of writer recognition, described in this paper, each person is represented by the set of modeled letters or symbols. The sketch of proposed method consists of three steps: first handwritten letter or symbol must be modeled by a curve, then compared with unknown letter and finally there is a decision of identification. Author recognition of handwriting and signature is based on the choice of key points and curve modeling. Reconstructed curve does not have to be smooth in the nodes because a writer does not think about smoothing during the handwriting. Curve interpolation in handwriting identification is not only a pure mathematical problem but important task in pattern recognition and artificial intelligence such as: biometric recognition [2-4], personalized handwriting recognition [5], automatic forensic document examination

[6,7], classification of ancient manuscripts [8]. Also writer recognition in monolingual handwritten texts is an extensive area of study and the methods independent from the language are well-seen. Proposed method represents language-independent and text-independent approach because it identifies the author via a single letter or symbol from the sample. This novel method is also applicable to short handwritten text.

Writer recognition methods in the recent years are going to various directions: writer recognition using multi-script handwritten texts [9], introduction of new features [10], combining different types of features [3], studying the sensitivity of character size on writer identification [11], investigating writer identification in multi-script environments [9], impact of ruling lines on writer identification [12], model perturbed handwriting [13], methods based on run-length features [14,3], the edge-direction and edge-hinge features [2], a combination of codebook and visual features extracted from chain code and polygonized representation of contours [15], the autoregressive coefficients [9], codebook and efficient code extraction methods [16], texture analysis with Gabor filters and extracting features [17], using Hidden Markov Model [18-20] or Gaussian Mixture Model [1]. But no method is dealing with writer identification via curve modeling or interpolation and points comparing as it is presented in this paper.

The author wants to approach a problem of curve interpolation [21-23] and shape modeling [24] by characteristic points in handwriting identification. Proposed method relies on nodes combination and functional modeling of curve points situated between the basic set of key points. The functions that are used in calculations represent whole family of elementary functions with inverse functions: polynomials, trigonometric, cyclometric, logarithmic, exponential and power function. These functions are treated as probability distribution functions in the range [0;1]. Nowadays methods apply mainly polynomial functions, for example Bernstein polynomials in Bezier curves, splines and NURBS [25]. But Bezier curves do not represent the interpolation method and cannot be used for example in signature and handwriting modeling with characteristic points (nodes). Numerical methods for data interpolation are based on polynomial or trigonometric functions, for example Lagrange, Newton, Aitken and Hermite methods. These methods have some weak sides [26] and are not sufficient for curve interpolation in the situations when the curve cannot be build by polynomials or trigonometric functions. Proposed 2D curve interpolation is the functional modeling via any elementary

functions and it helps us to fit the curve during handwriting identification.

This paper presents novel Probabilistic Nodes Combination (PNC) method of curve interpolation and takes up PNC method of two-dimensional curve modeling via the examples using the family of Hurwitz-Radon matrices (MHR method) [27], but not only (other nodes combinations). The method of PNC requires minimal assumptions: the only information about a curve is the set of at least two nodes. Proposed PNC method is applied in handwriting identification via different coefficients: polynomial, sinusoidal, cosinusoidal, tangent, cotangent, logarithmic, exponential, arc sin, arc cos, arc tan, arc cot or power. Function for PNC calculations is chosen individually at each modeling and it represents probability distribution function of parameter  $\alpha \in [0;1]$  for every point situated between two successive interpolation knots. PNC method uses nodes of the curve  $p_i = (x_i, y_i) \in \mathbf{R}^2$ ,  $i = 1, 2, \dots, n$ :

1. PNC needs 2 knots or more ( $n \geq 2$ );
2. If first node and last node are the same ( $p_1 = p_n$ ), then curve is closed (contour);
3. For more precise modeling knots ought to be settled at key points of the curve, for example local minimum or maximum and at least one node between two successive local extrema.

Condition 3 means for example the highest point of the curve in a particular orientation, convexity changing or curvature extrema. The goal of this paper is to answer the question: how to model a handwritten letter or symbol by a set of knots [28]?

## II. PROBABILISTIC INTERPOLATION

The method of PNC is computing points between two successive nodes of the curve: calculated points are interpolated and parameterized for real number  $\alpha \in [0;1]$  in the range of two successive nodes. PNC method uses the combinations of nodes  $p_1=(x_1, y_1)$ ,  $p_2=(x_2, y_2), \dots, p_n=(x_n, y_n)$  as  $h(p_1, p_2, \dots, p_m)$  and  $m = 1, 2, \dots, n$  to interpolate second coordinate  $y$  for first coordinate  $c = \alpha \cdot x_i + (1-\alpha) \cdot x_{i+1}$ ,  $i = 1, 2, \dots, n-1$ :

$$y(c) = \gamma \cdot y_i + (1-\gamma)y_{i+1} + \gamma(1-\gamma) \cdot h(p_1, p_2, \dots, p_m), \quad (1)$$

$\alpha \in [0;1]$ ,  $\gamma = F(\alpha) \in [0;1]$ .

Here are the examples of  $h$  computed for MHR method [29]:

$$h(p_1, p_2) = \frac{y_1}{x_1} x_2 + \frac{y_2}{x_2} x_1 \quad (2)$$

or

$$h(p_1, p_2, p_3, p_4) = \frac{1}{x_1^2 + x_3^2} (x_1 x_2 y_1 + x_2 x_3 y_3 + x_3 x_4 y_1 - x_1 x_4 y_3) + \frac{1}{x_2^2 + x_4^2} (x_1 x_2 y_2 + x_1 x_4 y_4 + x_3 x_4 y_2 - x_2 x_3 y_4)$$

The examples of other nodes combinations:

$$h(p_1, p_2) = \frac{y_1 x_2}{x_1 y_2} + \frac{y_2 x_1}{x_2 y_1}$$

or

$$h(p_1, p_2) = \frac{y_1 x_2}{y_2} + \frac{y_2 x_1}{y_1}$$

or

$$h(p_1, p_2) = x_1 y_1 + x_2 y_2$$

or

$$h(p_1, p_2) = x_1 x_2 + y_1 y_2$$

or

$$h(p_1, p_2, \dots, p_m) = 0$$

or

$$h(p_1) = x_1 y_1$$

or others. Nodes combination is chosen individually for each curve. Formula (1) represents curve parameterization as  $\alpha \in [0;1]$ :

$$x(\alpha) = \alpha \cdot x_i + (1-\alpha) \cdot x_{i+1}$$

and

$$y(\alpha) = F(\alpha) \cdot y_i + (1-F(\alpha))y_{i+1} + F(\alpha)(1-F(\alpha)) \cdot h(p_1, p_2, \dots, p_m)$$

,

$$y(\alpha) = F(\alpha) \cdot (y_i - y_{i+1} + (1-F(\alpha)) \cdot h(p_1, p_2, \dots, p_m)) + y_{i+1}$$

Proposed parameterization gives us the infinite number of possibilities for curve calculations (determined by choice of  $F$  and  $h$ ) as there is the infinite number of human signatures, handwritten letters and symbols. Nodes combination is the individual feature of each modeled curve (for example a handwritten letter or signature). Coefficient  $\gamma = F(\alpha)$  and nodes combination  $h$  are key factors in PNC curve interpolation and shape modeling.

### A. Interpolating Functions in PNC Modeling

Points settled between the nodes are computed using PNC method. Each real number  $c \in [a;b]$  is calculated by a convex combination  $c = \alpha \cdot a + (1-\alpha) \cdot b$  for

$$\alpha = \frac{b-c}{b-a} \in [0;1].$$

Key question is dealing with coefficient  $\gamma$  in (1). The simplest way of PNC calculation means  $h = 0$  and  $\gamma = \alpha$  (basic probability distribution). Then PNC represents a linear interpolation. MHR method [30] is not a linear interpolation. MHR [31] is the example of PNC modeling. Each interpolation requires specific distribution of parameter  $\alpha$  and  $\gamma$  (1) depends on parameter  $\alpha \in [0;1]$ :

$$\gamma = F(\alpha), \quad F: [0;1] \rightarrow [0;1], \quad F(0) = 0, \quad F(1) = 1$$

and  $F$  is strictly monotonic. Coefficient  $\gamma$  is calculated using different functions (polynomials, power functions, sine, cosine, tangent, cotangent, logarithm, exponent, arc sin, arc cos, arc tan or arc cot, also inverse functions) and choice of function is connected with initial requirements and curve specifications. Different values of coefficient  $\gamma$  are connected with applied functions  $F(\alpha)$ . These functions  $\gamma = F(\alpha)$  represent the examples of probability distribution functions for random variable  $\alpha \in [0;1]$  and real number  $s > 0$ :

$$\begin{aligned} \gamma &= \alpha^s, & \gamma &= \sin(\alpha^s \cdot \pi/2), & \gamma &= \sin^s(\alpha \cdot \pi/2), & \gamma &= 1 - \cos(\alpha^s \cdot \pi/2), \\ \gamma &= 1 - \cos^s(\alpha \cdot \pi/2), & \gamma &= \tan(\alpha^s \cdot \pi/4), & \gamma &= \tan^s(\alpha \cdot \pi/4), \\ \gamma &= \log_2(\alpha^s + 1), & \gamma &= \log_2^s(\alpha + 1), & \gamma &= (2^\alpha - 1)^s, \\ \gamma &= 2/\pi \cdot \arcsin(\alpha^s), & \gamma &= (2/\pi \cdot \arcsin \alpha)^s, & \gamma &= 1 - \end{aligned}$$

$$2/\pi \cdot \arccos(\alpha^s), \gamma=1-(2/\pi \cdot \arccos\alpha)^s, \gamma=4/\pi \cdot \arctan(\alpha^s), \\ \gamma=(4/\pi \cdot \arctan\alpha)^s, \gamma=\text{ctg}(\pi/2-\alpha^s \cdot \pi/4), \gamma=\text{ctg}^s(\pi/2- \\ \alpha \cdot \pi/4), \gamma=2-4/\pi \cdot \text{arccotg}(\alpha^s), \gamma=(2-4/\pi \cdot \text{arccotg}\alpha)^s.$$

Functions above, used in  $\gamma$  calculations, are strictly monotonic for random variable  $\alpha \in [0;1]$  as  $\gamma = F(\alpha)$  is probability distribution function. Also inverse functions  $F^{-1}(\alpha)$  are appropriate for  $\gamma$  calculations. Choice of function and value  $s$  depends on curve specifications and individual requirements. Considering nowadays used probability distribution functions for random variable  $\alpha \in [0;1]$  - one distribution is dealing with the range  $[0;1]$ : beta distribution. Probability density function  $f$  for random variable  $\alpha \in [0;1]$  is:

$$f(\alpha) = c \cdot \alpha^s \cdot (1-\alpha)^r, s \geq 0, r \geq 0. \quad (3)$$

When  $r = 0$  probability density function (3) represents  $f(\alpha) = c \cdot \alpha^s$  and then probability distribution function  $F$  is like  $f(\alpha) = 3\alpha^2$  and  $\gamma = \alpha^3$ . If  $s$  and  $r$  are positive integer numbers then  $\gamma$  is the polynomial, for example  $f(\alpha) = 6\alpha(1-\alpha)$  and  $\gamma = 3\alpha^2 - 2\alpha^3$ . Beta distribution gives us coefficient  $\gamma$  in (1) as polynomial because of interdependence between probability density  $f$  and distribution  $F$  functions:

$$f(\alpha) = F'(\alpha), F(\alpha) = \int_0^\alpha f(t) dt. \quad (4)$$

For example (4):  $f(\alpha) = \alpha \cdot e^\alpha$  and  $\gamma = F(\alpha) = (\alpha - 1)e^\alpha + 1$ .

What is very important in PNC method: two curves (for example a handwritten letter or signature) may have the same set of nodes but different  $h$  or  $\gamma$  results in different interpolations (Fig.6-14).

Algorithm of PNC interpolation and modeling (1) looks as follows:

**Step 1:** Choice of knots  $p_i$  at key points.

**Step 2:** Choice of nodes combination  $h(p_1, p_2, \dots, p_m)$ .

**Step 3:** Choice of distribution  $\gamma = F(\alpha)$ .

**Step 4:** Determining values of  $\alpha$ :  $\alpha = 0.1, 0.2 \dots 0.9$  (nine points) or  $0.01, 0.02 \dots 0.99$  (99 points) or others.

**Step 5:** The computations (1).

These five steps can be treated as the algorithm of PNC method of curve modeling and interpolation (1).

Curve interpolation has to implement the coefficients  $\gamma$ . Each strictly monotonic function  $F$  between points  $(0;0)$  and  $(1;1)$  can be used in PNC interpolation.

### III. HANDWRITING MODELING AND RECOGNITION

PNC method enables signature and handwriting recognition. This process of recognition consists of three parts:

1. Modeling – choice of nodes combination and probabilistic distribution function (1) for known signature or handwritten letters;
2. Unknown writer - choice of characteristic points (nodes) for unknown signature or handwritten word and the coefficients of points between nodes;
3. Decision of recognition - comparing the results of PNC interpolation for known models with coordinates of unknown text.

#### A. Modeling – the Basis of Patterns

Known letters or symbols ought to be modeled by the choice of nodes, determining specific nodes combination and characteristic probabilistic distribution function. For example a handwritten word or signature “rw” may look different for persons A, B or others. How to model “rw” for some persons via PNC method? Each model has to be described by the set of nodes for letters “r” and “w”, nodes combination  $h$  and a function  $\gamma = F(\alpha)$  for each letter. Less complicated models can take  $h(p_1, p_2, \dots, p_m) = 0$  and then the formula of interpolation (1) looks as follows:

$$y(c) = \gamma \cdot y_i + (1-\gamma)y_{i+1}.$$

It is linear interpolation for basic probability distribution ( $\gamma = \alpha$ ). How first letter “r” is modeled in three versions for nodes combination  $h = 0$  and  $\alpha = 0.1, 0.2 \dots 0.9$ ? Of course  $\alpha$  is a random variable and  $\alpha \in [0;1]$ .

##### Person A

Nodes (1;3), (3;1), (5;3), (7;3) and  $\gamma = F(\alpha) = \alpha^2$ :

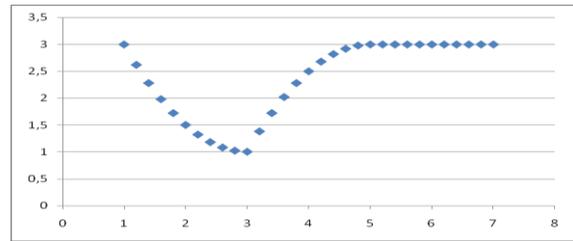


Fig. 1. PNC modeling for nine reconstructed points between nodes.

##### Person B

Nodes (1;3), (3;1), (5;3), (7;2) and  $\gamma = F(\alpha) = \alpha^2$ :

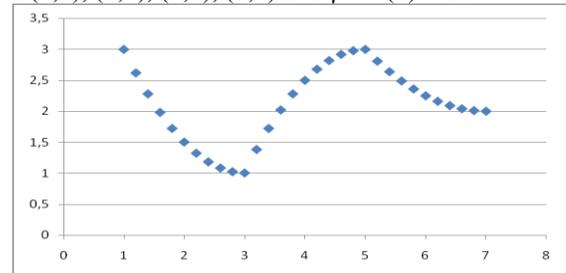


Fig. 2. PNC modeling of letter “r” with four nodes.

##### Person C

Nodes (1;3), (3;1), (5;3), (7;4) and  $\gamma = F(\alpha) = \alpha^3$ :

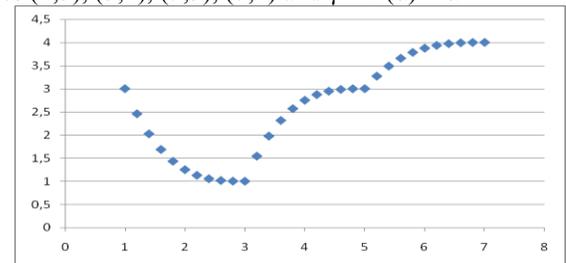


Fig. 3. PNC modeling of handwritten letter “r”.

These three versions of letter “r” (Fig.1-3) with nodes combination  $h = 0$  differ at fourth node and probability distribution functions  $\gamma = F(\alpha)$ . Much more possibilities of modeling are connected with a choice of nodes combination

$h(p_1, p_2, \dots, p_m)$ . MHR method [32] uses the combination (2) with good features because of orthogonal rows and columns at Hurwitz-Radon family of matrices:

$$h(p_i, p_{i+1}) = \frac{y_i}{x_i} x_{i+1} + \frac{y_{i+1}}{x_{i+1}} x_i$$

and then (1)

$$y(c) = \gamma \cdot y_i + (1-\gamma)y_{i+1} + \gamma(1-\gamma) \cdot h(p_i, p_{i+1})$$

Here are two examples of PNC modeling with MHR combination (2).

Person D

Nodes (1;3), (3;1), (5;3) and  $\gamma = F(\alpha) = \alpha^2$ :

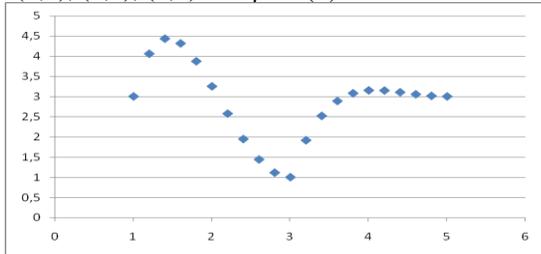


Fig. 4. PNC modeling of letter “r” with three nodes.

Person E

Nodes (1;3), (3;1), (5;3) and  $\gamma = F(\alpha) = \alpha^{1.5}$ :

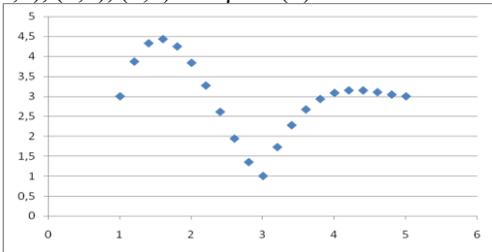


Fig. 5. PNC modeling of handwritten letter “r”.

Fig.1-5 show modeling of letter “r”. Now let us consider a letter “w” with nodes combination  $h = 0$ .

Person A

Nodes (2;2), (3;1), (4;2), (5;1), (6;2) and  $\gamma = F(\alpha) = (5^\alpha - 1)/4$ :

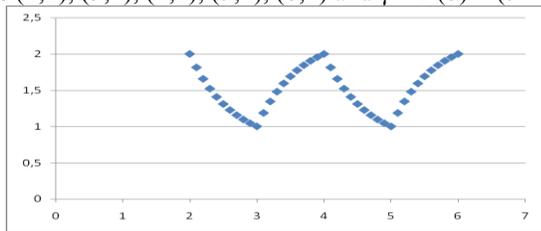


Fig. 6. PNC modeling for nine reconstructed points between nodes.

Person B

Nodes (2;2), (3;1), (4;2), (5;1), (6;2) and  $\gamma = F(\alpha) = \sin(\alpha \cdot \pi/2)$ :

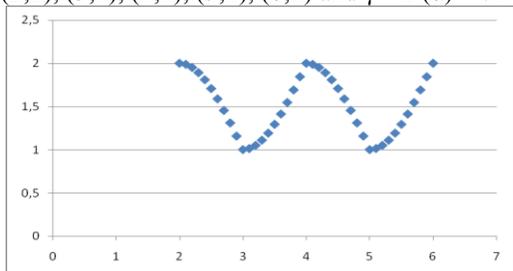


Fig. 7. PNC modeling of letter “w” with five nodes.

Person C

Nodes (2;2), (3;1), (4;2), (5;1), (6;2) and  $\gamma = F(\alpha) = \sin^{3.5}(\alpha \cdot \pi/2)$ :

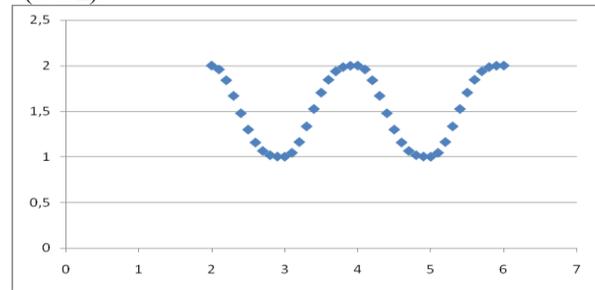


Fig. 8. PNC modeling of handwritten letter “w”.

These three versions of letter “w” (Fig.6-8) with nodes combination  $h = 0$  and the same nodes differ only at probability distribution functions  $\gamma = F(\alpha)$ . Fig.9 is the example of nodes combination  $h(2)$  from MHR method:

Person D

Nodes (2;2), (3;1), (4;1), (5;1), (6;2) and  $\gamma = F(\alpha) = 2^\alpha - 1$ :

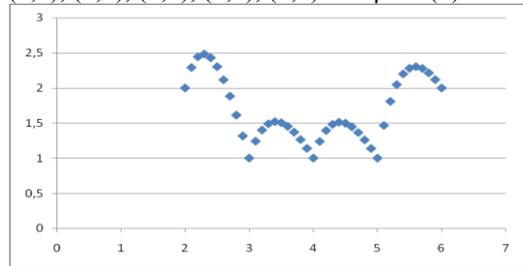


Fig. 9. PNC modeling for nine reconstructed points between nodes.

Examples above have one function  $\gamma = F(\alpha)$  and one combination  $h$  for all ranges between nodes. But it is possible to create a model with functions  $\gamma_i = F_i(\alpha)$  and combinations  $h_i$  individually for a range of nodes  $(p_i; p_{i+1})$ . It enables very precise modeling of handwritten symbol between each successive pair of nodes.

Each person has its own characteristic and individual handwritten letters, numbers or other marks. The range of coefficients  $x$  has to be the same for all models because of comparing appropriate coordinates  $y$ . Every letter is modeled by PNC via three factors: the set of nodes, probability distribution function  $\gamma = F(\alpha)$  and nodes combination  $h$ . These three factors are chosen individually for each letter, therefore this information about modeled letters seems to be enough for specific PNC curve interpolation, comparing and handwriting identification. Function  $\gamma$  is selected via the analysis of points between nodes and we may assume  $h = 0$  at the beginning. What is very important - PNC modeling is independent of the language or a kind of symbol (letters, numbers or others). One person may have several patterns for one handwritten letter. Summarize: every person has the basis of patterns for each handwritten letter or symbol, described by the set of nodes, probability distribution function  $\gamma = F(\alpha)$  and nodes combination  $h$ . Whole basis of patterns consists of models  $S_j$  for  $j = 0, 1, 2, 3, \dots, K$ .

### B. Unknown Author – Points of Handwritten Character

Choice of characteristic points (nodes) for unknown letter or handwritten symbol is a crucial factor in object recognition. The range of coefficients  $x$  has to be the same like the  $x$  range in the basis of patterns. Knots of the curve (opened or closed) ought to be settled at key points, for example local minimum or maximum (the highest point of the curve in a particular orientation), convexity changing or curvature maximum and at least one node between two successive key points. When the nodes are fixed, each coordinate of every chosen point on the curve  $(x_0^c, y_0^c), (x_1^c, y_1^c), \dots, (x_M^c, y_M^c)$  is accessible to be used for comparing with the models. Then probability distribution function  $\gamma = F(\alpha)$  and nodes combination  $h$  have to be taken from the basis of modeled letters to calculate appropriate second coordinates  $y_i^{(j)}$  of the pattern  $S_j$  for first coordinates  $x_i^c, i = 0, 1, \dots, M$ . After interpolation it is possible to compare given handwritten symbol with a letter in the basis of patterns.

### C. Recognition – the $W$ writer

Comparing the results of PNC interpolation for required second coordinates of a model in the basis of patterns with points on the curve  $(x_0^c, y_0^c), (x_1^c, y_1^c), \dots, (x_M^c, y_M^c)$ , we can say if the letter or symbol is written by person A, B or another. The comparison and decision of recognition [33] is done via minimal distance criterion. Curve points of unknown handwritten symbol are:  $(x_0^c, y_0^c), (x_1^c, y_1^c), \dots, (x_M^c, y_M^c)$ . The criterion of recognition for models  $S_j = \{(x_0^c, y_0^{(j)}), (x_1^c, y_1^{(j)}), \dots, (x_M^c, y_M^{(j)})\}, j=0, 1, 2, 3 \dots K$  is given as:

$$\sum_{i=0}^M |y_i^c - y_i^{(j)}| \rightarrow \min.$$

Minimal distance criterion helps us to fix a candidate for unknown writer as a person from the model  $S_j$ .

## IV. CONCLUSION

The method of Probabilistic Nodes Combination (PNC) enables interpolation and modeling of two-dimensional curves [34] using nodes combinations and different coefficients  $\gamma$ : polynomial, sinusoidal, cosinusoidal, tangent, cotangent, logarithmic, exponential, arc sin, arc cos, arc tan, arc cot or power function, also inverse functions. Function for  $\gamma$  calculations is chosen individually at each curve modeling and it is treated as probability distribution function:  $\gamma$  depends on initial requirements and curve specifications. PNC method leads to curve interpolation as handwriting or signature identification via discrete set of fixed knots. PNC makes possible the combination of two important problems: interpolation and modeling in a matter of writer identification. Main features of PNC method are:

- a) the smaller distance between knots the better;
- b) calculations for coordinates close to zero and near by extremum require more attention because of importance of these points;
- c) PNC interpolation develops a linear interpolation into other functions as probability distribution functions;

- d) PNC is a generalization of MHR method via different nodes combinations;
- e) interpolation of  $L$  points is connected with the computational cost of rank  $O(L)$  as in MHR method;
- f) nodes combination and coefficient  $\gamma$  are crucial in the process of curve probabilistic parameterization and interpolation: they are computed individually for a single curve.

Future works are going to: application of PNC method in signature and handwriting recognition, choice and features of nodes combinations and coefficient  $\gamma$ , implementation of PNC in computer vision and artificial intelligence: shape geometry, contour modelling, object recognition and curve parameterization.

## REFERENCES

- [1] Schlapbach, A., Bunke, H.: Off-line writer identification using Gaussian mixture models. In: International Conference on Pattern Recognition, pp. 992–995 (2006).
- [2] Bulacu, M., Schomaker, L.: Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Anal. Mach. Intell. 29 (4), 701–717 (2007).
- [3] Djeddi, C., Souici-Meslati, L.: A texture based approach for Arabic writer identification and verification. In: International Conference on Machine and Web Intelligence, pp. 115–120 (2010).
- [4] Djeddi, C., Souici-Meslati, L.: Artificial immune recognition system for Arabic writer identification. In: International Symposium on Innovation in Information and Communication Technology, pp. 159–165 (2011).
- [5] Nosary, A., Heutte, L., Paquet, T.: Unsupervised writer adaption applied to handwritten text recognition. Pattern Recogn. Lett. 37 (2), 385–388 (2004).
- [6] Van, E.M., Vuurpijl, L., Franke, K., Schomaker, L.: The WANDA measurement tool for forensic document examination. J. Forensic Doc. Exam. 16, 103–118 (2005).
- [7] Schomaker, L., Franke, K., Bulacu, M.: Using codebooks of fragmented connected-component contours in forensic and historic writer identification. Pattern Recogn. Lett. 28 (6), 719–727 (2007).
- [8] Siddiqi, I., Cloppet, F., Vincent, N.: Contour based features for the classification of ancient manuscripts. In: Conference of the International Graphonomics Society, pp. 226–229 (2009).
- [9] Garain, U., Paquet, T.: Off-line multi-script writer identification using AR coefficients. In: International Conference on Document Analysis and Recognition, pp. 991–995 (2009).
- [10] Bulacu, M., Schomaker, L., Brink, A.: Text-independent writer identification and verification on off-line Arabic handwriting. In: International Conference on Document Analysis and Recognition, pp. 769–773 (2007).
- [11] Ozaki, M., Adachi, Y., Ishii, N.: Examination of effects of character size on accuracy of writer recognition by new local arc method. In: International Conference on Knowledge-Based Intelligent Information and Engineering Systems, pp.1170–1175 (2006).
- [12] Chen, J., Lopresti, D., Kavallieratou, E.: The impact of ruling lines on writer identification. In: International Conference on Frontiers in Handwriting Recognition, pp. 439–444 (2010).
- [13] Chen, J., Cheng, W., Lopresti, D.: Using perturbed handwriting to support writer identification in the presence of severe data constraints. In: Document Recognition and Retrieval, pp. 1–10 (2011).
- [14] Galloway, M.M.: Texture analysis using gray level run lengths. Comput. Graphics Image Process. 4 (2), 172–179 (1975).
- [15] Siddiqi, I., Vincent, N.: Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. Pattern Recogn. Lett. 43 (11), 3853–3865 (2010).
- [16] Ghiasi, G., Safabakhsh, R.: Offline text-independent writer identification using codebook and efficient code extraction methods. Image and Vision Computing 31, 379–391 (2013).
- [17] Shahabinejad, F., Rahmati, M.: A new method for writer identification and verification based on Farsi/Arabic handwritten texts, Ninth

- International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 829–833 (2007).
- [18] Schlapbach, A., Bunke, H.: A writer identification and verification system using HMM based recognizers, *Pattern Anal. Appl.* 10, 33–43 (2007).
- [19] Schlapbach, A., Bunke, H.: Using HMM based recognizers for writer identification and verification, 9th Int. Workshop on Frontiers in Handwriting Recognition, pp. 167–172 (2004).
- [20] Marti, U.-V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition, *Int. J. Doc. Anal. Recognit.* 5, 39–46 (2002).
- [21] Collins II, G.W.: *Fundamental Numerical Methods and Data Analysis*. Case Western Reserve University (2003).
- [22] Chapra, S.C.: *Applied Numerical Methods*. McGraw-Hill (2012).
- [23] Ralston, A., Rabinowitz, P.: *A First Course in Numerical Analysis – Second Edition*. Dover Publications, New York (2001).
- [24] Zhang, D., Lu, G.: Review of Shape Representation and Description Techniques. *Pattern Recognition* 1(37), 1-19 (2004).
- [25] Schumaker, L.L.: *Spline Functions: Basic Theory*. Cambridge Mathematical Library (2007).
- [26] Dahlquist, G., Bjoerck, A.: *Numerical Methods*. Prentice Hall, New York (1974).
- [27] Jakóbczak, D.: 2D and 3D Image Modeling Using Hurwitz-Radon Matrices. *Polish Journal of Environmental Studies* 4A(16), 104-107 (2007).
- [28] Jakóbczak, D.: Shape Representation and Shape Coefficients via Method of Hurwitz-Radon Matrices. *Lecture Notes in Computer Science* 6374 (Computer Vision and Graphics: Proc. ICCVG 2010, Part I), Springer-Verlag Berlin Heidelberg, 411-419 (2010).
- [29] Jakóbczak, D.: Curve Interpolation Using Hurwitz-Radon Matrices. *Polish Journal of Environmental Studies* 3B(18), 126-130 (2009).
- [30] Jakóbczak, D.: Application of Hurwitz-Radon Matrices in Shape Representation. In: Banaszak, Z., Świć, A. (eds.) *Applied Computer Science: Modelling of Production Processes* 1(6), pp. 63-74. Lublin University of Technology Press, Lublin (2010).
- [31] Jakóbczak, D.: Object Modeling Using Method of Hurwitz-Radon Matrices of Rank  $k$ . In: Wolski, W., Borawski, M. (eds.) *Computer Graphics: Selected Issues*, pp. 79-90. University of Szczecin Press, Szczecin (2010).
- [32] Jakóbczak, D.: Implementation of Hurwitz-Radon Matrices in Shape Representation. In: Choraś, R.S. (ed.) *Advances in Intelligent and Soft Computing* 84, Image Processing and Communications: Challenges 2, pp. 39-50. Springer-Verlag, Berlin Heidelberg (2010).
- [33] Jakóbczak, D.: Object Recognition via Contour Points Reconstruction Using Hurwitz-Radon Matrices. In: Józefczyk, J., Orski, D. (eds.) *Knowledge-Based Intelligent System Advancements: Systemic and Cybernetic Approaches*, pp. 87-107. IGI Global, Hershey PA, USA (2011).
- [34] Jakóbczak, D.: Curve Parameterization and Curvature via Method of Hurwitz-Radon Matrices. *Image Processing & Communications- An International Journal* 1-2(16), 49-56 (2011).

Technical University of Koszalin, Poland and since October 2007 he has been an Assistant Professor in the Chair of Computer Science and Management in this department. His research interests connect mathematics with computer science and include computer vision, artificial intelligence, shape representation, curve interpolation, contour reconstruction and geometric modeling, numerical methods, probabilistic methods, game theory, operational research and discrete mathematics.



Dariusz Jacek Jakóbczak was born in Koszalin, Poland, on December 30, 1965. He graduated in mathematics (numerical methods and programming) from the University of Gdansk, Poland in 1990. He received the Ph.D. degree in 2007 in computer science from the Polish – Japanese Institute of Information Technology, Warsaw, Poland.

From 1991 to 1994 he was a civilian programmer in the High Military School in Koszalin. He was a teacher of mathematics and computer science in the Private Economic School in Koszalin from 1995 to 1999. Since March 1998 he has worked in the Department of Electronics and Computer Science,

# Effect of Precursor on Growth of MoS<sub>2</sub> Monolayer and Multilayer

Shraddha Ganorkar, Jungyoon Kim, Young Hwan Kim and Seong-II Kim\*

**Abstract**—The rise of two-dimensional (2D) material is one of the results of the successful efforts of researchers which laid the path to the new era of electronics. The size and dimensionality will reveal the new limits of electronic devices and applications. One of the most exciting materials is MoS<sub>2</sub>. In the last few years the MoS<sub>2</sub> has been studied extensively to understand its chemical kinematics and possible practical applications. Synthesis has been always a major issue as electronic devices need reproducibility along with similar properties for mass productions. Chemical vapor deposition (CVD) is one of the successful methods for 2D materials including graphene. Much of this research is still in its infancy, but this and other techniques will be developed and improved in the near future. Furthermore, there are various starting materials available for Mo and S source. The different source has different effects on the layers and morphology of MoS<sub>2</sub> films. In this work, we have extensively studied the CVD technique to grow few layers of MoS<sub>2</sub> with different starting materials and compare their results. We investigated the results of two precursors MoO<sub>3</sub> and MoCl<sub>5</sub>, show remarkable changes. The MoO<sub>3</sub> source gives a triangular shaped MoS<sub>2</sub> monolayer with Raman-shift  $\Delta k=21.5\text{ cm}^{-1}$  while that of MoCl<sub>5</sub> can achieve uniform MoS<sub>2</sub> without triangle. The photoluminescence spectra of monolayer MoS<sub>2</sub> grown from MoO<sub>3</sub> shows absorption peaks at 1.83 eV (675.73 nm) and 1.99 eV (621.78 nm). While bilayer MoS<sub>2</sub> film from MoCl<sub>5</sub> precursor shows absorption at 1.88 eV (657.44 nm) and 2.04 eV (605.10 nm). The film synthesized by MoCl<sub>5</sub> is more continuous and it would be a good choice for device applications. Eventually, we tried to explain the formation of continuous monolayer of MoS<sub>2</sub> without any triangle on the basis of chemical reaction formalism mostly like due to one step reaction process and formation of MoS<sub>2</sub> from gas phase to the solid phase.

**Keywords**—2D Materials, MoS<sub>2</sub> monolayer, CVD, Raman Spectra

## I. INTRODUCTION

**S**ILICON is the backbone of semiconductor industries from the last few decades. Owing to its remarkable properties like tunable bandgap (via doping) and switchable conductivity via magnetic or electric fields, temperature and even mechanical deformation; silicon behaves as ideal material for transistor or sensing device. This traditional semiconductor

This work was supported by Korea Institute of Science and Technology (2E25373). Shraddha Ganorkar, Jungyoon Kim, Young Hwan Kim and Seong-II Kim\* are with the Center for Nano Photonics, Korea Institute of Science and Technology, 5, Hwarangro 14-gil, Seongbuk-gu, Seoul, 136-791, South Korea (phone: 82-2-958-5737 ; fax: 82-2-958-5739 \*email:s-ikim@kist.re.kr).

now faces a daunting task. The new era of the high power and nano size are forcing electronic devices to reach new limits in fabrication. With the miniaturization of transistors (expected 14 nm in this year) the issues like short channel effects and defects densities will become harder to hold back. When the electronic industries seeking for a new material, the rise of 2D materials gave a light of hope.

Just over 10 years ago, isolation of graphene [1], the very first 2D carbon, a strong contender rose as a 2D device material. It has the extraordinary property of high electron mobility, excellent optical transmittance, thermal conductivity, large Young's modulus and chemical inertness [2]-[4]. These properties are highly sought after in a semiconducting industry. However, lack of a band gap and metallic behavior rules it out as a semiconductor. A strong bandgap engineering required for graphene for which it suffers its other properties. Nevertheless, graphene triggered a great deal of attention towards the 2D material. The search of 2D materials has thus grown to encompass other materials which exhibit similar properties to graphene and traditional semiconductors.

Among the new systems transition metal chalcogenides (TMCs) has shown the very similar properties demanded by electronic devices. One of the TMCs, molybdenum disulfide, MoS<sub>2</sub>, has a similar layer structure like graphene, the hexagons consist of covalently bonded Mo and S atoms. The Mo layer is covalently sandwiched between two S layers to give S-Mo-S layers which are stacked over each other by Van der Waals forces. It is an excellent candidate for device fabrication. It is well known that structures with nanometric dimensions have different electronic, chemical, optical and magnetic properties. Similar way, monolayer of MoS<sub>2</sub> has vastly different properties as compared to its bulk counterpart. Bulk MoS<sub>2</sub> has an indirect band gap of 1.29 eV while that of its monolayer has a large direct band gap of 1.8 eV [5]. The layered structure enables MoS<sub>2</sub> to have a tunable band gap based on the number of layers grown. Theoretical studies have also predicted the tunable band gap also possible with external electric field [6]. There are several sparkling properties of MoS<sub>2</sub> which makes it to be used in the potentials devices. MoS<sub>2</sub> has stiffness, resistant for braking, excellent mechanical properties [7], very high current density [8], high on/off ratio and electron mobility similar to silicon [9]. Furthermore, MoS<sub>2</sub> has strong fluorescence by virtue of its direct band gap. These properties coupled with its aforementioned tunable band gap, allow MoS<sub>2</sub> used in fabrication of ion of flexible electronics and optical sensing or

emitting devices at different optical frequencies and wavelength. Every new material needs a unique characterization tool. Since the 2D material has remarkably different properties from bulk, researchers have designed way to characterize and identify different types of monolayers. Raman spectroscopy and resonant Raman spectroscopy are one of the tools which allows one to distinguish and identify the material and the number of layers [10].

Synthesis of MoS<sub>2</sub> is still an open challenge for researchers. There are several methods to grow or fabricate monolayer MoS<sub>2</sub>. Some of which have been very successful. Top-down approaches like mechanical exfoliation and lithium intercalation assisted exfoliation which has less control on the thickness of layers. Chemical vapor deposition (CVD) has proven one of the successful technique for growing monolayer to few layers of 2D materials including graphene [11]-[17]. The number of layers critically depends on several factors in CVD like temperature, pressure, position of precursor and substrate, etc. Nevertheless, sources of Mo and S also play an important for deciding the morphology of the films. In this paper, we have investigated the CVD method. We used two different Mo precursors (MoO<sub>3</sub> and MoCl<sub>5</sub>) to grow MoS<sub>2</sub> films and compare their results. It has been observed that for practical application purposed the films grown with MoCl<sub>5</sub> can be a better choice for device fabrication.

## II. EXPERIMENT

The growth of MoS<sub>2</sub> monolayer and few layers were carried out in a home-built CVD furnace with 1 inch quartz tube. The precursor MoO<sub>3</sub> (Sigma-Aldrich 99.999%), MoCl<sub>5</sub> (Sigma-Aldrich 99.999%) and S (Sigma-Aldrich 99.999%) were used for MoS<sub>2</sub> synthesis. 300 nm SiO<sub>2</sub> on Si was used as substrate. The substrate was sonicated with trichloroethylene, acetone, methanol and DI water for 15min each. Since the two precursor need the different treatment the method is described below briefly.

### A. Synthesis with MoO<sub>3</sub> precursor

The MoO<sub>3</sub> was placed in a quartz boat at the center of the furnace with the substrate held upside down. The S was placed at upstream at 14 cm from the center of the furnace. Furnace was heated to 700 °C in 40 min with N<sub>2</sub> flow of 10 sccm. The temperature was held about 5 min and allow it to cool naturally to 500 °C followed by rapid cooling by opening furnace and flowing 500 sccm N<sub>2</sub>. With this method we able to grow MoS<sub>2</sub> monolayer and bilayer films.

### B. Synthesis with MoCl<sub>5</sub> Precursor

The MoCl<sub>5</sub> powder placed at the center of the furnace and the S powder is placed in the quartz crucible at the upstream of the furnace. The substrate was placed at downstream (next to Mo precursor) from 1-4cm away from the center of the furnace. The growth was carried out in 2 torr Ar atmosphere at 50 sccm flow. The furnace was heated to 800 °C in 30 min and held for 5 min followed by natural cooling to room temperature. With this method we could able to get MoS<sub>2</sub> mono, bi, tri and tetra layer films.

The films synthesized as above were characterized by optical microscope, Raman spectroscopy and photoluminescence spectroscopy (Uni-RAM 5500, UniNanoTech, Korea) with frequency-doubled Nd:YAG (532 nm) laser.

## III. RESULT AND DISCUSSION

Figure. 1 shows the optical images of MoS<sub>2</sub> layers grown using MoO<sub>3</sub>. The monolayer of MoS<sub>2</sub> can be clearly seen in Fig. 1 (a). Various size triangle-shaped MoS<sub>2</sub> grains can be clearly seen in the image. Some grains are merged to form star shapes and many arbitrary shapes which indicate the formation of

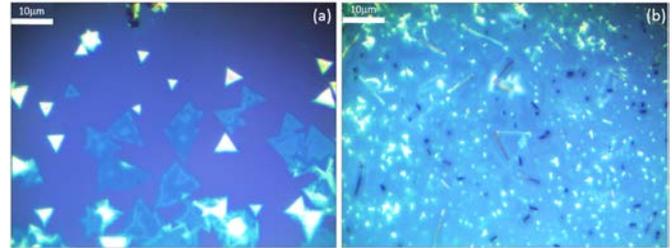


Fig. 1 Optical images of MoS<sub>2</sub> synthesized from MoO<sub>3</sub> precursor (a) Monolayer (1L) (b) Bilayer (2L)

continuous layer. Such an overlapping of grains turn into bilayer film, which can be seen from Fig. 1 (b). It is difficult to get very clean film using MoO<sub>3</sub> as source. We characterized the synthesized films using Raman spectroscopy. It is a powerful nondestructive characterization tool for MoS<sub>2</sub>. Raman spectra of the films grown using MoO<sub>3</sub> is shown in Fig. 2. Two characteristic Raman modes can be found in the Raman spectra. The off resonance first-order Raman active modes  $E_{2g}^1$  (387

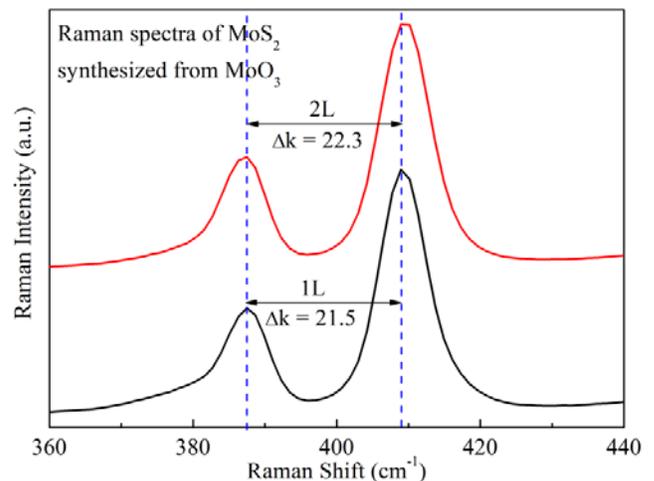


Fig. 2 Raman spectra of MoS<sub>2</sub> monolayer (1L) and bilayer (2L) synthesized from MoO<sub>3</sub>

cm<sup>-1</sup>) and  $A_{1g}$  (408 cm<sup>-1</sup>) are generally observed for bulk MoS<sub>2</sub>. The  $A_{1g}$  mode results from opposite vibration of two S atom with respect to Mo atom while  $A_{1g}$  mode arises from out of plane vibration of S atoms in the opposite direction [10]. These modes are closely related to number of layers. The Raman spectra of monolayer and bilayer MoS<sub>2</sub> grown from MoO<sub>3</sub> is

depicted in Fig. 2. The frequency difference ( $\Delta k$ ) between the Raman modes for monolayer is found to be  $\Delta k = 21.5 \text{ cm}^{-1}$  while that of for bilayer it is found to be  $\Delta k = 22.3 \text{ cm}^{-1}$ . The grown films show excellent optical quality. Photoluminescence (PL) for monolayer MoS<sub>2</sub> is shown in Fig. 3. Two PL peaks can be observed around 675 nm (1.83 eV) and 621 nm (1.99 eV) corresponding to the A<sub>1</sub> and B<sub>1</sub>, respectively for direct excitonic transitions with the energy split from the valence band spin-orbital coupling [18]. It is difficult to grow controlled multilayer with MoO<sub>3</sub> source

The second precursor was MoCl<sub>5</sub>. Though the method of synthesis was CVD but MoCl<sub>5</sub> need the different growth conditions to grow MoS<sub>2</sub> multilayers. The optical images of

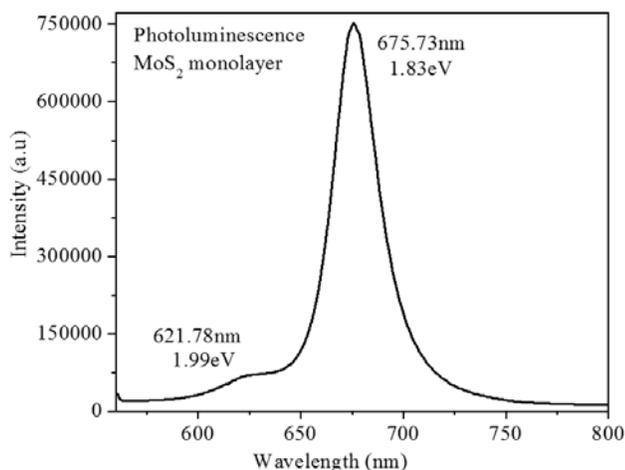


Fig. 3 Photoluminescence of MoS<sub>2</sub> monolayer grown by MoO<sub>3</sub>

MoS<sub>2</sub> grown by MoCl<sub>5</sub> source are depicted in Fig 4 (a) monolayer 1L and Fig. 4 (b) bilayer (2L). One can notice the there are no triangle observed like MoO<sub>3</sub> growth. In contrast to earlier Mo source, these films are found to be very uniform. We have successfully grown the various layers by tailoring the

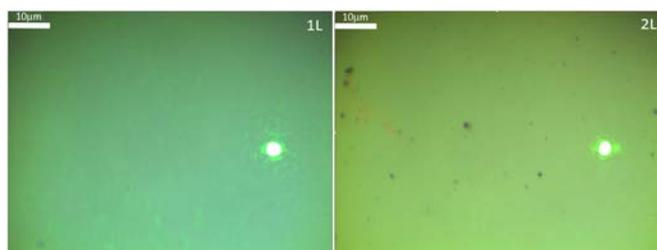


Fig. 4 Optical images of MoS<sub>2</sub> synthesized from MoCl<sub>5</sub> precursor (a) Monolayer (1L) (b) Bilayer (2L)

substrate position (1 to 4 cm from the source). We obtained the thicker film for the larger distance between the source and substrate. The Raman spectra of MoS<sub>2</sub> multilayers are shown in Fig. 5. One can observe the significant increase in the Raman frequency difference  $\Delta k$  with increases of MoS<sub>2</sub> thickness from monolayer to tetralayer. A systematic correlation is found between the Raman modes and number of layers. PL spectra of bilayer MoS<sub>2</sub> grown from MoCl<sub>5</sub> source is presented in Fig. 6. PL peaks A<sub>1</sub> (657.44 nm, 1.88 eV) and B<sub>1</sub> (605.10 nm, 2.04 eV)

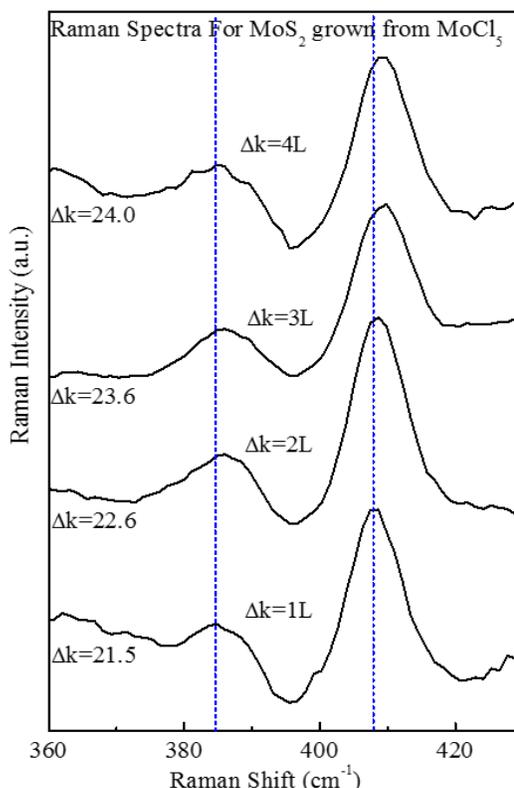


Fig. 5 Raman spectra of MoS<sub>2</sub> monolayer (1L), bilayer (2L), trilayer (3L) and tetralayer (4L) synthesized from MoCl<sub>5</sub>

observed at lower wavelength than monolayer, as expected due to increases in the band gap of bilayer film. One can notice drastic decrease in PL peak intensity as compared to monolayer. This justifies the evolution of bandgap with an

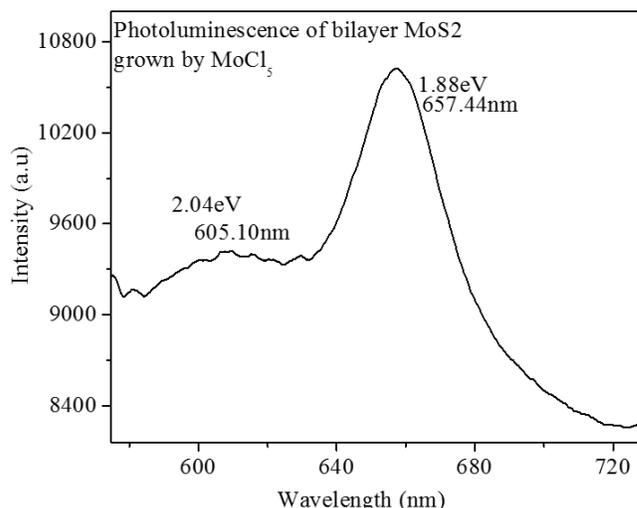
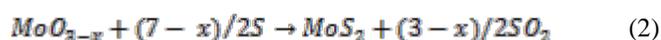


Fig. 6 Photoluminescence spectra of MoS<sub>2</sub> bilayer film grown by MoCl<sub>5</sub> source

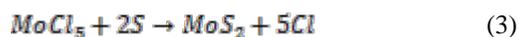
increase in the number of layers. It is well known fact that there is no observable PL spectra in bulk MoS<sub>2</sub> is due to local field effect while strong PL spectra of monolayer shows that luminescence quantum efficiency is much higher in monolayer than in multilayer and bulk [18].

From the above studies once can see that in CVD process the

source materials play an important role for deciding the morphology of the film. MoS<sub>2</sub> monolayer grows in different geometrical shapes triangular to hexagonal when MoO<sub>3</sub> is used as source material. There are several factors in the CVD process which decide the shapes of monolayer MoS<sub>2</sub>. A possible explanation for shape evolution can be given by a principle of crystal growth in which the shape change of domains is attributed to local changes in the source ratio (Mo:S) as well as its influence on the kinetic growth dynamics of edges [19]. The mechanism of reduction of MoO<sub>3</sub> to MoS<sub>2</sub> in the absence of H is still not known. It is believed that the reaction between MoO<sub>3</sub> and S involves stepwise reduction of Mo<sup>VI</sup> in MoO<sub>3</sub> to Mo<sup>IV</sup> in MoS<sub>2</sub>. This transition is supposed to involve reduction and sulfurization. A postulated stepwise process is given by the following equations [20], [21]



In these steps there is possibility of formation of oxysulfide (MoOS<sub>2</sub>), which is a composite of MoS<sub>2</sub> and MoO<sub>(3-x)</sub>. MoO<sub>2</sub> is one of the most stable intermediate in this process. Substituting  $x=1$  in equation (1) realized the formation of MoO<sub>2</sub>. Where as in case of MoCl<sub>5</sub> in presence of excess S is possibly one step process. The chemical reaction between MoCl<sub>5</sub> and S is not clearly reported as per best of our knowledge. The possible process can be given as follows.



Furthermore the growth of MoS<sub>2</sub> multilayers with MoCl<sub>5</sub> source is believed to be “self-limiting” process. It is most postulated that formation MoS<sub>2</sub> in gas phase followed by its diffusion onto receiving substrate and further precipitation to MoS<sub>2</sub> solid phase [15]. This indeed supports for our assumption of one step chemical reaction of MoCl<sub>5</sub> and S. From this scenario, we can conclude that the one step chemical reaction is the key point to get uniform monolayer of MoS<sub>2</sub> without formation of triangles. Since in this particular process there is no need of nucleation, which is the main reason for formation of geometric shapes (mostly triangular) of MoS<sub>2</sub> monolayer. In device application of MoS<sub>2</sub> monolayer the geometric shapes can create the issues related to grain bounties. Hence MoCl<sub>5</sub> or similar sources can be a good choice as a precursor for MoS<sub>2</sub> deposition.

#### REFERENCES

- [1] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zang, S. V. Dubonos, I. V. Grigorieva, A. A. Firsov, “Electric field effect in atomically thin carbon films,” *Science*, vol. 306, pp. 666-669, Oct, 2004 DOI: 10.1126/science.1102896
- [2] K. S. Novoselov, V. I. Fal’ko, L. Colombo, P. R. Gellert, M. G. Schwab, K. Kim, “A roadmap for graphene,” *NATURE*, vol. 490, pp. 192-200, Oct, 2012 DOI:10.1038/nature11458
- [3] Y. Zhu, S. Murali, W. Cai, X. Li, J. W. Suk, J. R. Potts, R. S. Ruoff, “Graphene and graphene oxide: synthesis, properties, and applications,” *Adv. Mater.*, vol. 22, pp. 3906-3924, Jun, 2010 DOI:10.1002/adma.201001068
- [4] X. Song, J. Hu, H. Zeng, “Two-dimensional semiconductors: recent progress and future perspectives,” *J. Mater. Chem. C*, vol. 1, pp. 2952-2969, Jan, 2013 DOI: 10.1039/C3TC00710C
- [5] K. F. Mak, C. Lee, J. Hone, J. Shan, T. F. Heinz, “Atomically thin MoS<sub>2</sub>: a new direct-gap semiconductor,” *Physical Rev. Lett.*, vol. 105, pp. 136805(1-4), Sep, 2010 DOI: 10.1103/PhysRevLett.105.136805
- [6] A. Ramasubramaniam, D. Naveh, E. Towe, “Tunable band gaps in bilayer transition-metal dichalcogenides,” *Physical Review B*, vol. 84, pp. 205325(1-10), Nov, 2011 DOI: 10.1103/PhysRevB.84.205325
- [7] S. Bertolazzi, J. Brivio, A. Kis. “Stretching and breaking of ultrathin MoS<sub>2</sub>,” *ACS Nano*, vol5, pp. 9703-9709, Nov, 2011 DOI: 10.1021/nn203879f
- [8] D. Lembke, A. Kis, “Breakdown of high-performance monolayer MoS<sub>2</sub> transistors,” *ACS Nano*, vol. 6, pp. 10070-10075, Oct, 2012 DOI:10.1021/nn303772b
- [9] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, A. Kis. “Single-layer MoS<sub>2</sub> transistors,” *Nat. Nanotechnol.*, vol. 6, pp. 147-150, Jan, 2011 DOI:10.1038/nnano.2010.279
- [10] H. Li, Q. Zhang, C. C. R. Yap, B. K. Tay, T. H. T. Edwin, A. Olivier, D. Baillargeat, “From bulk to monolayer MoS<sub>2</sub> : evolution of Raman scattering,” *Adv. Funct. Mater.*, vol. 22, pp. 1385-1390, Jan, 2012 DOI: 10.1002/adfm.201102111
- [11] J. Mann, D. Sun, Q. Ma, J. R. Chen, E. Preciado, T. Ohta, B. Diaconescu, K. Yamaguchi, T. Tran, M. Wurch, et al. “Facile growth of monolayer MoS<sub>2</sub> film areas on SiO<sub>2</sub>,” *Eur. Phys. J. B*, vol. 86, pp. 226(1-4), May, 2013 DOI: 10.1140/epjb/e2013-31011-y
- [12] Y. H. Lee, X. Q. Zhang, W. Zhang, M. T. Chang, C. T. Lin, K. D. Chang, Y. C. Yu, J. T. W. Wang, C. S. Chang, L. J. Li, et al, “Synthesis of large-area MoS<sub>2</sub> atomic layers with chemical vapor deposition,” *Adv. Mater.*, vol. 24, pp. 2320-2325, May, 2012 DOI: 10.1002/adma.201104798
- [13] S. Najmaei, Z. Liu, W. Zhou, X. Zou, G. Shi, S. Lei, B. I. Yakobson, J. C. Idrobo, P. M. Ajayan, J. Lou, “Vapor phase growth and grain boundary structure of molybdenum disulphide atomic layers,” *Nat. Mater.*, vol. 12, pp. 754-759, Jun, 2013 DOI:10.1038/nmat3673
- [14] Y. C. Lin, W. Zhang, J. K. Huang, K. K. Liu, Y. H. Lee, C. T. Liang, C. W. Chu, L. J. Li, “Wafer-scale MoS<sub>2</sub> thin layers prepared by MoO<sub>3</sub> sulfurization,” *Nanoscale*, vol. 4, pp. 6637-6641, Aug, 2012 DOI: 10.1039/c2nr31833d
- [15] Y. Yu, C. Li, Y. Liu, L. Su, Y. Zhang, L. Cao, “Controlled scalable synthesis of uniform, high-quality monolayer and few-layer MoS<sub>2</sub> films,” *Sci. Rep.*, vol. 3, pp. 1866, May, 2013 DOI:10.1038/srep01866
- [16] S. Balendhran, J. Z. Ou, M. Bhaskaran, S. Sriram, S. Ippolito, Z. Vasic, E. Kats, S. Bhargava, S. Zhuiykov, K. Kalantar-Zadeh, “Atomically thin layers of MoS<sub>2</sub> via a two-step thermal evaporation-exfoliation method,” *Nanoscale*, vol. 4, pp. 461-466, Nov, 2012 DOI:10.1039/C1NR10803D
- [17] Q. Ji, Y. Zhang, T. Gao, Y. Zhang, D. Ma, M. Liu, Y. Chen, X. Qiao, P. H. Tan, M. Kan, “Epitaxial monolayer MoS<sub>2</sub> on mica with novel photoluminescence,” *Nano Lett.*, vol. 13, pp. 3870-3877, Jul, 2013 DOI: 10.1021/nl401938t
- [18] A. Splendiani, L. Sun, Y. Zhang, T. Li, J. Kim, C. Y. Chim, G. Galli and F. Wang “Emerging photoluminescence in monolayer MoS<sub>2</sub>” *Nano Lett.*, vol 10 (4), pp. 1271-1275 Mar 2010 DOI: 10.1021/nl903868w
- [19] S. Wang, Y. Rong, Y. Fan, M. Pacios, H. Bhaskaran, K. He and J. H. Warner, “Shape evolution of monolayer MoS<sub>2</sub> crystals growth by chemical vapor deposition,” *Chem. Mater.* vol 26, pp. 6371-6379 Nov 2014 DOI: 10.1021/cm5025662
- [20] Y. D. Li and X. L. Li, “Formation of MoS<sub>2</sub> inorganic fullerenes (Ifs) by reaction of MoO<sub>3</sub> nanobelts and S,” *Chem. Eur. J.*, vol 9, pp. 2726- 2731 Jun 2003 DOI: 10.1002/chem.200204635
- [21] B. Li, S. Yang, N. huo, Y. Li, J. Yang, R. Li, C. Fan and F. Lu, “Growth of large area few-layer or monolayer MoS<sub>2</sub> from controllable MoO<sub>3</sub> nanowire nuclei,” *RSC Adv.*, , vol 4, pp. 26407, Apr 2014 DOI: 10.1039/c4ra01632g

# Photonic Crystal Cavities for Optical Signal Processing

Nikolay L. Kazanskiy, Pavel G. Serafimovich

**Abstract**— We describe and numerically investigate an all-optical temporal integrator and differentiator based on photonic crystal nanobeam cavities. We show that an array of photonic crystal cavities enables high-order temporal integration. The model of two-component nanocavity with possibility of vertical electrical pumping is also described.

**Keywords**— Photonic crystal cavities, optical data processing.

## I. INTRODUCTION

All-optical fully integrated on-chip computing components will increase the speed of information processing by several orders of magnitude. Moreover, such components enable the processing of not only real, but also complex values. In this regard, it is important to implement the basic computing operations optically. In recent years, all-optical integrators and differentiators based on Bragg gratings [1] and ring resonators [2] have been proposed. Such elements can be used in both digital and analog signal processing. Among the digital signal processing applications are the use of optical integrators and differentiators as pulse counters and ultrafast memory elements [3]. Analog signal processing applications include all-optical solution of differential equations of various orders [4].

Integrators and differentiators based on Bragg gratings are a few millimeters in size. Integrators and differentiators based on ring resonators are more compact. Their size is on the order of tens of micrometers on the chip plane. In this paper we describe and study numerically the most compact optical integrators and differentiators based on photonic crystal (PC) cavities [5].

The model of two-component nanocavity is also described. In this model, the minimum details of the structure are found only in the periodic component of the resonator. The

advantages of such a structure include a promising way to construct an electrically pumped photonic cavity, the ease of introducing non-linear optical materials in the area of the nanocavity, the possibility of formation of the desired energy distribution in the far zone, and the possibility to develop dynamic systems based on nanocavities.

## II. TEMPORAL INTEGRATION OF OPTICAL SIGNALS

Fig. 1 shows a scheme of the coupled-resonator optical waveguide (CROW). The variable  $a_i$ ,  $i = [1, N]$  is the complex amplitude of the resonant mode in the  $i$ -th resonator;  $\kappa_{i-1}$  and  $\kappa_i$ ,  $i = [1, N]$ , are the left and right coupling coefficients of the  $i$ -th resonator, respectively;  $r_i$ ,  $i = [1, N]$  is the energy loss of the  $i$ -th resonator to the exterior space; and  $p_{in}$ ,  $p_{rf}$ , and  $p_{tr}$  are the amplitude of the input, reflected, and transmitted fields, respectively.

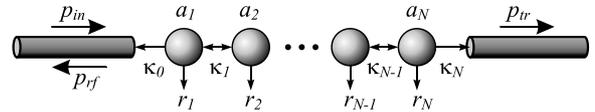


Fig. 1 Scheme of coupled-resonator optical waveguide.

Consider an array in which the resonators have identical resonant frequencies. Then, according to the temporal coupled-mode theory, the transmission function of the system can be written as

$$T_N(s) \equiv \frac{p_{tr}}{p_{in}} = -\frac{2(-\mathbf{i})^{N-1} \sqrt{\kappa_0 \kappa_N \kappa_1 \kappa_2 \cdots \kappa_{N-1}}}{\det(\mathbf{M})} \quad (1)$$

where  $\mathbf{M}$  is the corresponding tridiagonal matrix [6],  $p_{tr} = -\mathbf{i} \sqrt{2\kappa_N} a_N = -2\sqrt{\kappa_0 \kappa_N} [\mathbf{M}^{-1}]_{N,1} p_{in}$ , and  $\det(\mathbf{M})$  is the determinant of the matrix  $\mathbf{M}$ .

For  $N = 1$ , Eq. (1) is reduced to the form

$$T_1(s) = -\frac{2\kappa_0}{s + 2\kappa_0} \quad (2)$$

Hereafter, for simplicity, we neglect the losses to the exterior space.

This work was supported by the RFBR grants 13-07-97002, 13-07-13166, 14-07-97008, 14-07-97009, Ministry of Education and Science of the Russian Federation, and Programs ONIT RAS NN. 2 and 5.

Nikolay L. Kazanskiy is with the Image Processing Systems Institute of the Russian Academy of Sciences and Samara State Aerospace University (National Research University), 151, Molodogvardeyskaya st, Samara 443001 Russia (e-mail: kazansky@smr.ru).

Pavel G. Serafimovich is with the Image Processing Systems Institute of the Russian Academy of Sciences and Samara State Aerospace University (National Research University), 151, Molodogvardeyskaya st, Samara 443001 Russia (phone: 792-774-43563; e-mail: serp@smr.ru).

Let us consider how accurately Eq. (2) approximates the integrator of the first order. The polarized electric field with envelope  $P_{in}(t)$  can be written as

$$\begin{aligned} E(x,t) &= P_{in}(t - x/v_g) \exp(\mathbf{i}m_0x - \mathbf{i}\omega_0t) = \\ &= \int_{-\infty}^{\infty} R(\omega - \omega_0) \exp(\mathbf{i}m(\omega)x - \mathbf{i}\omega t) d\omega, \end{aligned} \quad (3)$$

where  $R(\omega)$  is the envelope spectrum signal,  $m(\omega)$  is the wave number [ $m_0 = m(\omega_0)$ ], and  $v_g$  is the group velocity.

A linear system described by the complex transfer function (TF)  $H(\omega)$  converts the envelope of the input pulse [Eq. (3)] to

$$P_{tr}(t) = \int_{-\infty}^{\infty} R(\omega) H(\omega) \exp(\mathbf{i}\omega t) d\omega = P_{in}(t) * h(t),$$

where the symbol  $*$  denotes the convolution operation, and  $h(t)$  is the spectrum of the TF  $H(\omega)$ .

The impulse response of a linear system with the TF  $T_1(s)$  is

$$h_1(t) = -\kappa_0 \exp(-\kappa_0 t) u(t) \quad (5)$$

where  $u(t)$  is the Heaviside step function.

Substituting Eq. (5) into Eq. (4), we obtain an expression for the envelope of the output pulse:

$$P_{tr}(t) = -\kappa_0 \int_{-\infty}^t P_{in}(T) \exp(-\mathbf{i}\kappa_0(t-T)) dT \quad (6)$$

The right side of this equation expresses the integral of the input pulse envelope with exponential weight.

Fig. 2 shows the result of integration of the envelope of an optical pulse with a duration of 100 ps by resonators with  $Q$ -factors of  $3 \times 10^4$  and  $5 \times 10^4$ . The  $Q$ -factor is related to  $\kappa_0$  by the ratio  $Q = \omega_0 / (4\kappa_0)$ . The figure shows that the higher  $Q$ -factor of the resonator is, the more slowly the integrated signal envelope decays. For resonators with  $Q$ -factors of  $3 \times 10^4$  and  $5 \times 10^4$  we estimate an integration time window (defined as the decay time required to reach 80% of the maximum intensity) of 12.5 ps and 19.5 ps, respectively.

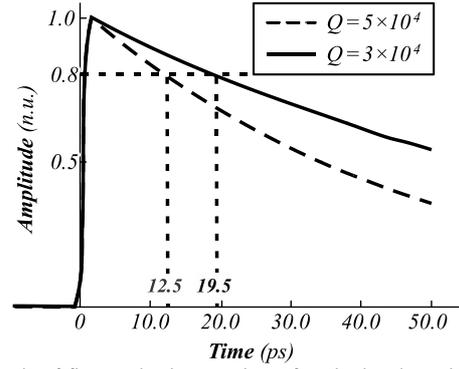


Fig. 2 Result of first-order integration of optical pulse with duration of 1 ps by resonators with  $Q$ -factors of  $3 \times 10^4$  and  $5 \times 10^4$ .

For  $N = 2$ , Eq. (1) can be written as

$$T_2(s) = \frac{\mathbf{i}2\kappa_0\kappa_1}{(s + \kappa_0)^2 + \kappa_1^2} \quad (7)$$

Let us calculate the parameters of the particular PC nanobeam cavity that integrates the optical signal. Compared with the resonators in the two-dimensional PC layer [5], PC nanobeam cavities have a smaller area and are naturally integrated into the waveguide geometry of the chip.

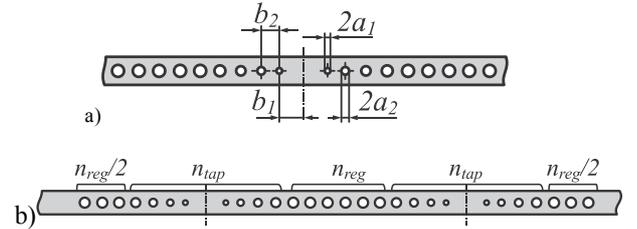


Fig. 3 Schemes of (a) PC nanobeam cavity and (b) array of two such cavities.

Fig. 3(a) illustrates a possible embodiment of a resonator based on a PC nanobeam. The decreasing radius of holes in the tapering region forms a defect in which the resonant mode is excited. An array of two PC resonators is shown in Fig. 3(b), where  $n_{tap}$  is the number of holes in the tapering region. The coupling value between resonators in the array is determined by  $n_{reg}$ , the number of holes with the maximum radius between defects. It can be shown [6] that for two adjacent resonators with quality factors  $Q_1$  and  $Q_2$ , the coupling coefficient is

$$\kappa = \frac{\omega_0}{4\sqrt{Q_1Q_2}} = \frac{\omega_0}{4Q_0} a^{-n_{reg}} \quad (8)$$

where  $Q_0$  is the  $Q$ -factor of the resonator containing only the hole defect zone ( $n_{reg} = 0$ ), and  $\omega_0$  is the resonant frequency corresponding to the Bragg wavelength. The value of  $a$  can be

approximated from the calculation of the  $Q$ -factor of a single resonator with different values of  $n_{reg}$ .

The resonance cavity characteristics were computed using the parallel 3D finite-difference time-domain method [7]. The waveguide in our simulations has a width of 490 nm and a height of 220 nm. It is composed of silicon and deposited on silica substrate. Air-filled holes in the regular part of the waveguide have a radius of 100 nm and are spaced 330 nm apart. The lattice parameters and the radii of the holes near the defect ( $n_{top}=12$ ) are following in nm:  $a_1=40$ ,  $b_1=255$ ,  $a_2=55$ ,  $b_2=350$ ,  $a_3=65$ ,  $b_3=365$ ,  $a_4=75$ ,  $b_4=375$ ,  $a_5=85$ ,  $b_5=385$ ,  $a_6=95$ ,  $b_6=395$ . These parameters demonstrate the existence of an energy bandgap for transverse electric polarization in the waveguide.

If the number of left and right holes is equal to  $n_{reg}/2$ , as shown in Fig. 2(b), then the following condition holds:

$$\kappa_0 = \kappa_N = \kappa_j, j = 1, N - 1 \quad (9)$$

Fig. 4(a) shows the result of integration of the first derivative of a Gaussian pulse with a duration of 150 fs. The  $Q$ -factor of the resonator is  $3.6 \times 10^4$ , and  $n_{reg} = 6$ . The result of integration of the second derivative of a pulse with the same duration by second-order integrator is shown in Fig. 4(b). The second-order integrator consists of two resonators, as shown in Fig. 3(b). The coupling coefficients between the resonators in each integrator are given by Eq. (8).

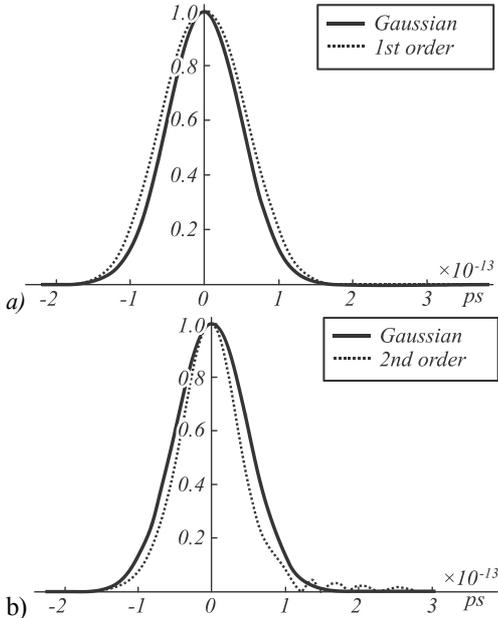


Fig. 4 Results of integration of corresponding derivatives of Gaussian pulse with duration of 150 fs by (a) one PC resonator, (b) an array of two PC resonators.

### III. TEMPORAL DIFFERENTIATION OF OPTICAL SIGNALS

The use of a PC resonator as an optical differentiator is described in [8]. The function of the complex reflection of such a resonator can be written as

$$H_{df}(s) = -\frac{s}{s + \kappa_0} \quad (10)$$

The value of  $\omega_0$  is zero for the ideal differentiator function [ $H_{idf}(s) = -s$ ]. To illustrate the above-described approach to the implementation of the resonator-aided differentiators, a resonator based on the ridge photonic-crystal waveguides (RPhCW) is designed in this section.

In the RPhC waveguide, the total internal reflection prevents light from propagating in the transverse directions, whereas the reflection of light in the longitudinal direction of the nanoresonator is enabled by the photonic crystal.

To design an optical resonator with high  $Q$ -factor ( $Q = \omega_0 \tau / 2$ ), a two-component structure is used (top inset in Fig. 5). First, these are photonic-crystal mirrors with identical equidistant holes in the waveguides.  $Q_w$  ( $Q_w = \omega_0 \tau_w / 2$ ) can be increased by increasing the number of these holes. As a result of calculation, the waveguide has the width  $w = 710$  nm and the height  $h = 230$  nm. The air-filled waveguide holes of radius of  $R = 90$  nm are spaced 330-nm ( $a = 330$  nm) apart. Such geometric parameters enable the emergence of a bandgap for the TE-wave in the waveguide. The waveguide is made of silicon ( $n = 3.46$ ) and is placed in air.

The second component of the structure is a transition zone between the PhC waveguide and the resonance cavity. This zone is intended to reduce the scattering losses in the resonator.  $Q_r$  ( $Q_r = \omega_0 \tau_r / 2$ ) can be increased by minimizing the spatial Fourier harmonics of the cavity mode inside the waveguide lightcone. This achieved by creating a Gaussian field attenuation with quadratic tapering the filling fraction  $f$  ( $f = \pi R^2 / (aw)$ ). The first 10 photonic crystal mirror segments (counted from the center) have filling fraction varying from 0.2 to 0.1 and, correspondingly, the holes radii  $R$  varying from 125 to 90 nm.

The pulse source of light is found in the left (or right) part of the waveguide in Fig. 5. The pulse central frequency is  $1.2 \cdot 10^6$  GHz (wavelength- 1.55  $\mu\text{m}$ ), corresponding to the position of the bandgap of the PhC waveguide.

The resonance cavity characteristics were computed using the parallel FDTD method. Absorbing layers were placed at the boundaries of the three-dimensional region under computation. The computational grid resolution was chosen so as to attain a converging solution.

$Q_r$  was calculated through incrementing the number of the PhC layers in the resonator's mirrors. The plot in Fig. 5 depicts the resonator's  $Q$ -factor as a function of the number of the layers. At the saturation point, when  $N=10$ , we obtain the

value  $Q_r \approx Q = 4 \times 10^5$ . If photonic-crystal layers in the cavity mirrors are absent, then  $Q = 1.15 \times 10^3$ . The differentiating resonator is supposed to have a relatively low  $Q$ -factor when compared with the resonators used for other applications. As a result, a fairly wide spectral range of short pulses can be processed. A small value of the  $Q$ -factor leads to a low-level reflected signal at the resonance frequency.

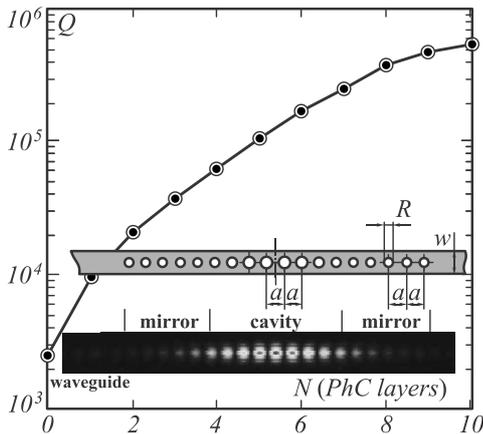


Fig. 5 The  $Q$ -factor as a function of the number of the PhC layers in the resonator's mirrors. The top inset depicts the designed nanoresonator's structure. The bottom inset shows the energy density distribution within the resonator.

Figure 6 shows the differentiation results for a pulse with the envelope defined by the function  $\exp(-x^2/(2\sigma^2))$ . The values of  $Q$  and  $Q_r$  that were used for the results of Fig. 6 correspond to Fig. 5 ( $Q=1.15 \times 10^3$ ,  $Q_r=4 \times 10^5$ ). The device length is  $20 \times 0.33 = 6.6 \mu\text{m}$  (20 holes in the cavity region, the holes in the mirrors region are absent). The result of the pulse differentiation performed by means of the nanoresonator designed is presented in Fig. 6 (a). The root-mean-square (r.m.s.) deviation of the analytically derived (solid line) curve from the numerically simulated (dotted line) curve is 46%. Shown in Fig. 6 (b) are similar results for a  $\sim 50$ -ps pulse. In this case, the r.m.s. deviation is 29%. Shown in Fig. 6 (c) are the results derived for a  $\sim 100$ -ps pulse. The r.m.s. deviation is 4%.

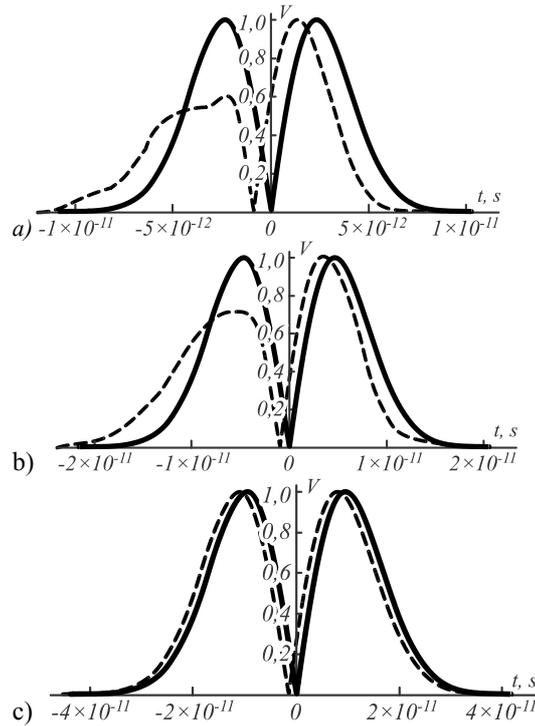


Fig. 6 The differentiation results for a pulse with the envelope defined by  $\exp(-x^2/(2\sigma^2))$ . (a) The differentiation result for the  $\sim 25$ -ps pulse obtained with the resonator designed (r.m.s.=46%). (b) Differentiation results for a  $\sim 50$ -ps pulse (r.m.s.=29%). (c) Differentiation results for a  $\sim 100$ -ps pulse (r.m.s.=4%).

#### IV. TWO-COMPONENT CAVITY STRUCTURE

Most of the existing technologies in use to create high- $Q$  PC nanocavities suggest fine-tuning of the resonance chamber geometry by changing the parameters of the photonic crystal. Such parameters may be, for example, the radius of the hole in the photonic crystal period and/or the hole periodic spacing. To simplify the solutions for the problems cited, the authors theoretically investigated two-component PC cavity. The first component of such a cavity is a periodic structure on the basis of a PC nanobeam. Compared with the two-dimensional structure on the basis of a PC slab, the area of PC nanobeam is smaller and is naturally integrated into the waveguide geometry of connections on a chip. The second component is a fragment of a complementary material having an area of several lattice constant of the PC. The shape and size of the fragment were determined from the given parameters of PC cavity. While combining the two components, a defect forms in the resulting nanostructure. The resonant mode of the corresponding frequency can be excited in this defect.

The proposed approach to creating two-component PC cavities is illustrated through the structure shown in Fig. 7. The first component of the resonator was a PC nanobeam. The nanobeam was made of silicon and was placed on a silica substrate. The holes in the nanobeam were of the same radius; they were equidistant from each other and filled with air. PC

nanobeam parameters are given in the legend to Fig. 7. With these parameters, PC band gap was created for TE- dominant polarization radiation in the wavelength range of 1.4-1.7  $\mu\text{m}$ . The second component of the nanocavity was an elliptical piece of silicon, placed on a silica substrate.

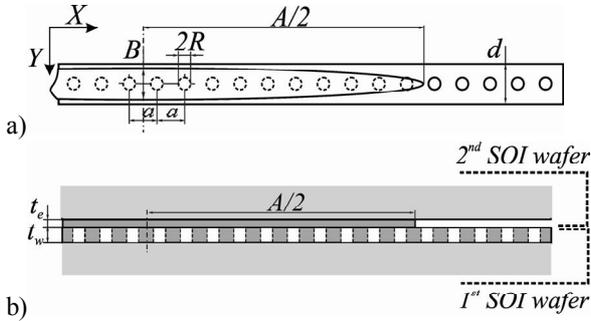


Fig. 7 The geometry of the resonator calculated by (a) top view, (b) a side view. PC nanobeam ( $n = 3.46$ ) lies on the substrate ( $n = 1.45$ ). PC nanobeam width is  $d = 0.5 \mu\text{m}$ , thickness  $t_w = 0.26 \mu\text{m}$ . Circular holes have a radius of  $R = 75 \text{ nm}$  and filled with air, distance between holes  $a = 0.34 \mu\text{m}$ . The elliptical shape (ellipse parameters  $A$  and  $B$ ) ( $n = 3.46$ ) lies on the substrate ( $n = 1.45$ ). Thickness of ellipse  $t_e = 100 \text{ nm}$ .

Therefore, both the cavity's components have a structure of Silicon on Insulator (SOI) wafers. The surface roughness of silicon wafers can be as low as several hundreds picometres at 1 - 300  $\mu\text{m}$  length scales. This permits to tightly combine two silicon surfaces of wafers as shown on Fig. 7.

In subsequent calculations, the thicknesses of the PC nanobeam and elliptical defect were assumed to be 260 nm and 100 nm, respectively. These thicknesses produced an optimal FF change in the cavity. Increasing the thickness of the nanobeam necessitates increase in the thickness of the elliptical defect.

The resonance cavity characteristics were computed using the parallel FDTD method. In particular, the cavity ( $Q = 3.05 \times 10^4$ ) was calculated with the parameters of the ellipse  $A = 6.8 \mu\text{m}$  (20 holes below the ellipse) and  $B = 0.5$ . To achieve a high-Q nanocavity, five additional holes were placed in the PC nanobeam at both ends of the ellipse. Thus, the total length of the cavity was  $(20 + 5 \times 2) \times 0.34 = 10.2 \mu\text{m}$ . Fig. 8a shows the distribution of  $H_z$  in the vertical plane passing through the axis of the nanobeam. There is some vertical asymmetry of the resonance mode due to flow of energy in the elliptical defect. Fig. 8b shows the distribution in the horizontal plane, just above the elliptical silicon fragment (in silica).  $H_z$  values along the intersection line of these two planes are represented by the dotted line in the graph of Fig. 8c, and the values directly below the PC nanobeam (in silica) by the dashed line on the same chart. The solid line represents the function  $\cos(\pi x/a) \exp(-\sigma x^2)$  with  $\sigma = 0.23$ ,  $a = 0.34$  microns. Good agreement between the distributions of  $H_z$  and an analytic function demonstrates that

the shape of the resonant mode's envelope is Gaussian. Assuming a linear dependence of  $\gamma$  on  $x$ , the relation  $\gamma(x) = a/\pi \int \sigma dx \approx x/40$  can be obtained. In [9], quadratic tapering of PC nanobeam width was used to form the defect. In the paper [9], for a nanocavity with a length of 60 periods of the PC, the relation  $\gamma(x) \approx x/120$  was implemented. Thus, it can be concluded that the two techniques used in creating a defect are almost equivalent. The nanocavity with an elliptical defect is 3 times shorter than the one with variable nanobeam width. Accordingly, the rate of change  $\gamma$  in the nanocavity with an elliptical defect is three times faster [10].

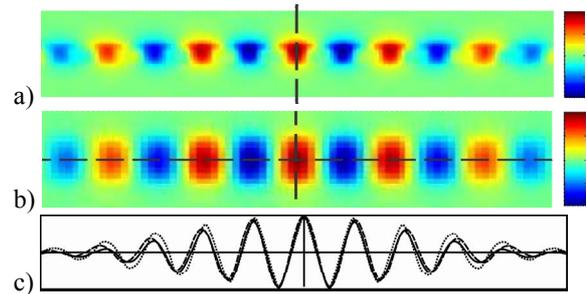


Fig. 8 (a) the distribution of  $H_z$  in the vertical plane passing through the axis of the waveguide, (b) the distribution of  $H_z$  in the horizontal plane just above the elliptical fragment (in quartz), (c) the dotted line -  $H_z$  values along the line of intersection of the planes (a) and (b), the dashed line -  $H_z$  values just below PC nanobeam (in quartz), the solid line - function  $\cos(\pi x/a) \exp(-\sigma x^2)$  for  $\sigma = 0.23$ ,  $a = 0.34 \mu\text{m}$ .

The two-component nanocavity proposed in this paper has in our opinion two main advantages when compared to existing solutions. First, the proposed structure allows for the development of an integrated on-chip light source with vertical electrical pumping. Integrated on-chip light-emitting diodes with a laterally doped p-i-n structure, based on the nanobeam photonic crystal cavity, were demonstrated recently. Electron beam lithography steps can be used to implant N- and P- type dopants to the first and second components of the structure, respectively. Fig. 9 shows an example of geometry for P- type and N- type doping regions. Such geometry permits to focus current flow to the active region of the cavity, thereby, in comparison with lateral electrical pumping, improving efficiency and reducing threshold. The P- type parts that adjoin the elliptical defect have a small intersection with the resonance mode. Therefore, Q-factor suffered low degradation, especially in case of  $B > d$  (Fig. 7a).

Using hybrid metal/photonic-crystal nanocavities is another possible approach to realize vertical electrical pumping. The two-component structure of the cavity assumes additional flexibility in choice of electrical current pathways. Although Q-factor in this case can be reduced to several

hundreds, this could be enough for development of an optical amplifier integrated on-chip.

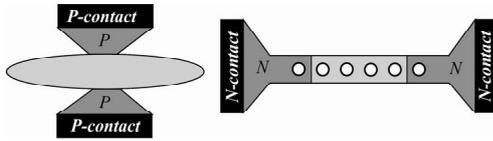


Fig. 9 An example of geometry (not to scale) for P- type (left) and N- type (right) doping regions.

The second advantage is that the creation of nanocavities with nonlinear properties is simplified. The supplementary component of the structure can be used to bring a nonlinear or optically active material directly into the nanocavity region.

## V. CONCLUSIONS

The authors describe and numerically investigate an all-optical temporal integrator and differentiator based on photonic crystal nanobeam cavities. We show that an array of photonic crystal cavities enables high-order temporal integration. This integrator is more compact than any of those previously suggested. Its dimensions depend linearly on the order of integration. The model of two-component nanocavity with possibility of electrical pumping is also described. In this model, the minimum details of the structure are found only in the periodic component of the resonator. The advantages of such a structure include a promising way to construct an electrically pumped photonic cavity, the ease of introducing non-linear optical materials in the area of the nanocavity, the possibility of formation of the desired energy distribution in the far zone, and the possibility to construct dynamic systems based on nanocavities.

## REFERENCES

- [1] Ngo, N. Q., "Design of an optical temporal integrator based on a phase-shifted Bragg grating in transmission," *Opt. Lett.* 32(20), 3020–3022 (2007).
- [2] Ferrera, M. et al, "On-chip CMOS-compatible all-optical integrator," *Nat. Commun.* 1, 1 (2010).
- [3] Ding, Y., Zhang, X., Zhang, X. and Huang, D., "Active microring optical integrator associated with electroabsorption modulators for high speed low light power loadable and erasable optical memory unit," *Opt. Express* 17(15), 12835–12848 (2009).
- [4] Slavik, R., et al, "Photonic temporal integrator for all-optical computing," *Opt. Express*. 16(22), 18202-18214 (2008).
- [5] Akahane, Y., Asano, T., Song, B.-S. and Noda, S., "Fine-tuned high-Q photonic-crystal nanocavity," *Opt. Express* 13(4), 1202–1214 (2005).
- [6] Liu, H. C. and Yariv, A., "Designing coupled-resonator optical waveguides based on high-Q tapered grating-defect resonators," *Opt. Express* 20(8) 9249-9263 (2012).
- [7] Kazanskiy N. L., Serafimovich P. G., "Coupled-resonator optical waveguides for temporal integration of optical signals," *Opt. Express* 22(11) 14004-14013 (2014).
- [8] Kazanskiy N.L., Serafimovich P.G., Khonina S.N. "Use of photonic crystal cavities for temporal differentiation of optical signals," *Opt. Letters* 38(7) 1149-1151 (2013).

- [9] Quan Q. and Loncar M., "Deterministic design of high Q, small mode volume photonic crystal nanobeam cavities," *Opt. Express* 19(5) 18529–18542 (2011).
- [10] Serafimovich P.G., Kazanskiy N.L., Khonina S.N. "Two-component cavity based on a regular photonic crystal nanobeam," *Applied Optics* 52(23) 5830-5834 (2013).

# A stateless Key Management Technique for Protection of Sensitive Data at Proxy Level for SQL based Databases using NIST Recommended SP800-132

*Kurra Mallaiiah*

Osmania University, Hyderabad  
[km\\_mallaiiah@yahoo.com](mailto:km_mallaiiah@yahoo.com)

*Prof. S Ramachandram*

Osmania University, Hyderabad  
[schandram@gmail.com](mailto:schandram@gmail.com)

**Abstract:** *Protecting confidentiality of organizational sensitive data in outsourced databases is continuously raising security concerns. The activities of malicious administrator and malicious software attacks on database server while data in use are alarming security concerns in the outsourced database servers. Therefore data needs to be protected in its complete life cycle (at rest, in transition and while in use) to realize better confidentiality for the outsourced databases. Data must be in encrypted form in the premises of database service providers even at the time of computations on data i.e. all the possible computation performed on encrypted data without decrypting in the premises of service providers. CryptDB is a practical and provable solution from MIT towards providing the confidentiality for relational databases by supporting the computations on encrypted data at database server side. For shared data, among different users, CryptDB use application's access control policy at the level of SQL queries in the proxy and each principal encryption key is encrypted with his password and for shared sensitive data owner's key is encrypted separately with all authorized users key and stored in the proxy and principal key is accessed with password key chaining technique. In this paper, we are proposing a stateless Key management technique in contrast to state full key management of CryptDB at the proxy level to protect sensitive data of users along with shared data for out sourced SQL databases using NIST recommended SP800-132. This approach eliminates the storing of user assigned encryption keys and also speak for relation keys in the proxy.*

**Keywords:** *Key management, Proxy, database security, cloud computing*

## 1. Introduction

Key management in cryptography is the organization of tasks concerned with protecting, storing, backing up and administration of encryption keys. High sensitive data losses and regulatory compliance requirements have spurred a dramatic increase in the use of encryption in the enterprise. The problem is that a single enterprise might use several different and possibly incompatible encryption algorithms to protect sensitive data, resulting in huge number of encryption keys and these keys must be protected and efficiently used. According to Verizon PCI

Compliance Report (PCIR) [3] about 42 percent of organizations have trouble implementing a proper encryption key management strategy to keep information safe. Expert says, proper encryption key management is becoming more important than encryption itself.

Encryption keys represent "the keys to the kingdom," if someone has access to the encryption key; they have access to the most sensitive data in your organization's encrypted data. Proper encryption key management is a requirement for PCI-DSS compliance [4]. Even auditors are scrutinizing how organizations manage encryption keys [5]. Regulatory compliances are recommending that storage of keys along with the data should be avoided. Therefore, key management playing an important role in protection of organizational sensitive data. Now days, the trend is moving towards outsourcing of organizational data to cloud data centers rapidly. However, at the same time apprehensive about usage of cloud data centers particularly key management for shared data among different users. Databases are one of the most compromised assets according to the 2014 Verizon Data Breach Report. The reason databases are targeted is quite simple; databases are at the heart of any organization, storing customer records and other confidential business data. Organizations are not protecting these assets well enough. According to IDC, less than 5% of the \$27 billion spent on security products directly addressed data center security. In cloud databases, organizational data is going through various stages. Organizational sensitive data is stored in the cloud database i.e. data is residing in cloud databases (Data at rest). Data is moving from client to server and vice-versa (data in transition). In addition, data is being manipulated at database server side (data in use) for attending to user queries. Data remains in three stages such as data at rest, data in transition and data in use. The protections of first two stages are well addressed but data while in use protection is not addressed adequately. The protection of data at rest is achieved by encrypting before storing onto database. For this there are many encryption algorithms are available notable among them are

symmetric encryption algorithms (AES, Blowfish, 3DES) and asymmetric encryption algorithms (RSA, ECC). The protection of data in transition can be achieved with available encryption, hashing and authentication mechanisms. SSL and TLS can be used to protect data while in transition; however, there was an attack while using the SSL 3.0 recently (POODLE). Data while in use also needs to be protected. Computations need to be performed on encrypted data without decryption at server side. This ensures confidentiality in database servers while data in use. There are some solutions, which allow computations directly on encryption data. Fully homomorphic encryption [6] is one such solution where it allows randomly computations on encrypted data. This technique is prohibitively slow, for real time applications (database intensive applications) may not suitable. Mylar [7] also supports searching on encrypted document in the context of multiuser and multi keys. However, this technique uses public key encryption mechanism, which is considered quite expensive compared with symmetric key technique and need to share part of the public key with the server and it targeted for only searching the encrypted documents. CryptDB [1] supports operation on encrypted data at server side using the SQL aware encryption algorithms and all the required encryption keys are stored in CryptDB. It is a kind of gateway solution, where required security on the organizational sensitive data is applied before sending to the cloud database server for storing. In this technique, a small change is required in application side and no change is required in database server side. As per Garter, these kinds of solutions (Cloud Security Broker Architecture) are going to evolve more and more. In CryptDB, key management is based on the chaining of passwords to encryption keys and keys which are used for encryption of user data and shared data keys as per speak for relation are stored in the proxy. In this paper, we are proposing a stateless key management (Not storing any keys in the proxy) mechanism based on the NIST recommended SP800-132 key derivation function contrast to state full technique of CryptDB key management (encryption and decryption keys are stored in the proxy) for user sensitive data protection at the proxy level for SQL based databases.

There are broadly four approaches to protect the sensitive data stored in the database servers in the cloud computing environment.

**Approach-1**

This approach demands necessary changes to the database server to achieve the desirable security to the enterprise sensitive data in the database server at source code level. This approach requires modifying the security logic at server side per requirement basic at source code level, which is a cumbersome activity and generally legacy systems does not support this approach and for every

application, requirement database server is required to change.

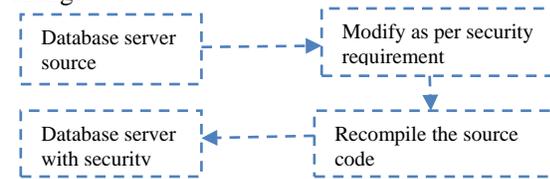


Fig 1: Change in database server at source code level.

**Approach-2**

As per the approach one can think of encrypting the data before hosting onto the cloud, bring complete encrypted data from server to client, and decrypt the encrypted data at client side as and when data is required for any analysis or manipulations. Entire database needs to be stored back on the cloud database, whenever there are any modifications in the database after re-encryption. In this approach, entire database needs to fetch from server to client for OLTP and OLAP transactions, which is an inefficient and resource consuming way of doing. SPORC [8] and Depot [9] are based on this approach

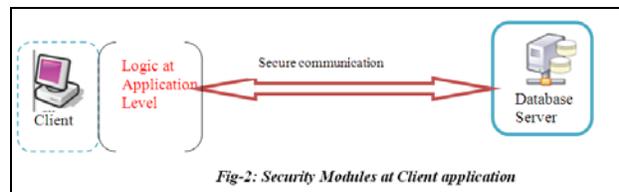


Fig-2: Security Modules at Client application

**Approach-3**

In this approach, logic may be incorporated in the server side, where encryption and decryption takes place before write and read to and from database. This approach requires enterprise sensitive data needs to be decrypted at server side for all operations; moreover, the key management is with server side. Therefore, the enterprise sensitive data completely in the hands of third party hence it could be exploited in many ways by malicious administrators or hackers. SUNDR [10], Oracle’s TDE [11] are based on this approach.

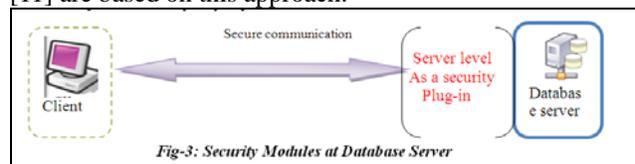


Fig-3: Security Modules at Database Server

**Approach-4**

In this approach, logic is set at intermediate level i.e. at proxy level to provide the security to the enterprise sensitive data. Security measures are applied on sensitive data before storing onto the cloud database server such a way that operations are directly possible on encrypted data itself in the database server. As per this approach

either at client/application side or server side need not necessary to change/incorporate any new logic or modify or change existing logic. This approach is completely transparent to the applications and servers. This approach is very promising per SQL based databases and for NoSQL databases. Based on this approach, a CryptDB technique is presented for protecting the confidentiality of sensitive data for SQL backed database applications. Ciphercloud gateway [12] and Navajo Systems [13] also based on this approach.

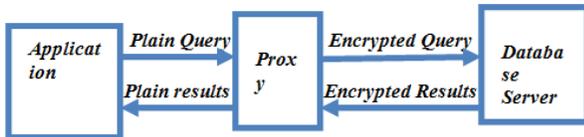


Fig 4: Security at intermediate (proxy) level

## 2. CryptDB Key management

Key management in CryptDB is based on chaining of password to the encryption key, which is used for encryption data of that user. Encryption key of each principal user is encrypted with password of that user and stored in the proxy. When user login to the application with password, this password is used to decrypt the encryption keys stored in the proxy corresponding to that user. The concept of key chaining to the password ensures that data is accessed using only owner's encryption key and it is possible only to owner of that data when he login into application using password. For shared data, CryptDB maintains speak for relation for that data. Example, user 'A' of type u speaks for user 'B' of type v, meaning that 'A' has access to all keys that 'B' has access to. Example, If User B speaks for principal A (as a result of some SPEAKS FOR annotation), then principal A's key is encrypted using principal B's key, and stored as a row in the special access keys table in the database. This allows principal B to gain access to principal A's key. The keys, used for encryption of data of each user are a combination of a symmetric key and a public-private key pair. In the common case, CryptDB uses the symmetric key of a user to encrypt any data and other user's keys accessible to this user. However, this is not always possible, if some principal is not currently online. For example, Suppose Bob sends message X to Alice, but Alice is not online. This means that CryptDB does not have access to Alice's key, so it will not be able to encrypt message X's key with Alice's symmetric key. In this case, CryptDB looks up the public key of the principal (Alice) in another table, public keys, and encrypts message X's key using Alice's public key. When Alice logs in, she will be able to use the secret key part of her key to decrypt the key for message X (and re-encrypt it under her symmetric key for future use). In Brief the following procedure explain the key management of

CryptDB user own data and shared data among different users.

### Procedure for Key Generation and usage in CryptDB

1. Each user is assigned a random number and used as an encryption key (MK)
2. Passwords are used to encrypt the keys (MK) and stored in a Key Table for all the users.
3. Encryption keys(MK) are decrypted using passwords of corresponding users
4. While encrypting the sensitive data, ENC\_FOR and SPEAK\_FOR relations are maintained
5. To maintain SPEAK\_FOR relation for a given sensitive shared data, the owner of the (ENC\_FOR) sensitive data KEY is encrypted with all the users keys based on the SPEAK\_FOR relation separately and stored.
6. When user is not logged-in, but wanted to share the sensitive data then public-private key of un-logged users are used.

Under the scenario of application and proxy compromise, CryptDB is protecting the data of un-logged users of the application. However, logged user's information CryptDB does not protect, this is because the keys are used for encrypting the user data is encrypted using user password and stored in the proxy system. It is only possible to the proxy to decrypt this key when user login. User application provides the password to the proxy when user logins into application. Proxy uses this password to decrypt the user assigned key that is used for encrypting the user sensitive data and other user's keys as per SPEAK\_FOR relation. Un-logged user's sensitive data are protected because proxy does not hold the passwords of these users therefore unable to decrypt the keys, even proxy compromised any adversary does not able to decrypt the keys.

Now, let us consider the following situations where CryptDB key management may be in unsafe state.

- i. Let say database portion of proxy has been hacked where encryption keys are stored in tables by encrypting with user passwords. If adversary grip the user password while transmission from application to proxy then adversary uses this key to decrypt the stored keys in the proxy database. Therefore, storing encryption keys in the proxy may lead to compromise user sensitive data stored in the database server even for the un-logged users.
- ii. Why should store the user's keys which are randomly generated by encrypting with password of respective user in the proxy? An adversary may use crypto analysis to break the encryption and can able to retrieve the original encryption key from the proxy of the users who are not logged in.
- iii. Suppose, user login password is lost then proxy unable to retrieve the original key because encryption key can only be decrypted with user password used for encrypting the user data.

- iv. Attacks on user machines such as XSS, If application user name and password is stolen using the Cross Site Scripting (XSS) from web applications of users then encryption keys of that user will be compromised and subsequently shared data of the other users; because, for shared data principal user key is encrypted with others users keys.
- v. Integrity of user query results is not verified in the proxy.

In this paper, we have proposed key management techniques to handle some of these situations more efficiently for protection of data at the proxy level. The technique is based on the NIST recommended SP800-132, deriving an encryption key based on user password at proxy level dynamically without storing any encryption and decryption keys in the proxy.

### 3. NIST Special Publication 800-132 Recommendation for Password-Based Key Derivation

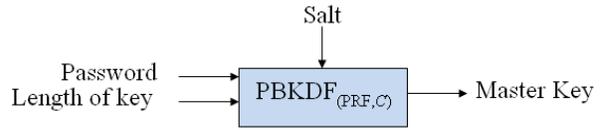
Outsourced organizational sensitive data's confidentiality mostly based on the encryption techniques. Encryption secrecy is complete based on the keys used for encryption, therefore, usage, protection and distribution of the keys plays a major role in maintaining the confidentiality of the organizational sensitive data.

The unpredictability of cryptographic keys is essential for the security of cryptographic based applications. There may be few situations with respect applications in that security of data is based on user passwords and this may be the only input required from the users who are eligible to access the sensitive data. Due to the low entropy and possibly poor randomness of those passwords, they are not suitable to be used directly as cryptographic keys [5]. However, NIST has recommended for Password-Based Key Derivation. The following section describes procedure to derive the encryption keys using the user provided password.

A password or a passphrase is a string of characters that is usually chosen by a user. Passwords are often used to authenticate a user in order to allow access to a resource. Since most user-chosen passwords have low entropy and weak randomness properties these passwords shall not be used directly as cryptographic keys. However, in certain applications, such as protecting data in storage devices, the password may be the only secret information that is available to the cryptographic algorithm that protects the data. KDFs are deterministic algorithms that are used to derive cryptographic keying material from a secret value, such as a password. Each PBKDF in the family is defined by the choice of a Pseudorandom Function (PRF) and a fixed iteration count, denoted as  $C$ . The input to an execution of PBKDF includes a password, denoted as  $P$ , a salt, denoted as  $S$ , and an indication of the desired length of the MK in bits, denoted as  $kLen$ .

Symbolically:

$$MK = \text{PBKDF}_{(\text{PRF}, C)}(P, S, kLen).$$



The  $kLen$  value shall be at least 112 bits in length. A minimum iteration count of 1,000 is recommended. For especially critical keys, or for very powerful systems or systems where user-perceived performance is not critical, an iteration count of 10,000,000 may be appropriate. The following is an algorithmic procedure to generate the master key using password [2]

Input:  $P$  password  
 $S$  Salt  
 $C$  Iteration Count

$kLen$  :Length of MK in bits; at most  $(2^{32} - 1) \times hLen$

**Parameter:** PRF HMAC with an **approved** hash function

$hlen$  Digest size of the hash function

**Output:**  $mk$  Master key

**Algorithm:**

**If**  $(kLen > (2^{32} - 1) \times hLen)$

**Return** an error indicator and stop;

$len = \lceil kLen/hLen \rceil$  ;

$r = kLen - (len - 1) \times hLen$  ;

**For**  $i = 1$  to  $len$

$T_i = 0$ ;

$U_0 = S \parallel \text{Int}(i)$ ;

**For**  $j = 1$  to  $C$

$U_j = \text{HMAC}(P, U_{j-1})$

$T_i = T_i \oplus U_j$

**Return**  $mk = T_1 \parallel T_2 \parallel \dots \parallel T_{len} <0\dots r-1>$

The KDF are deterministic, therefore generated Key will be always same for the given password, salt and count. Hence, this key can be used for encryption and decryption of sensitive data by deriving using PBKDF function dynamically in the proxy. This technique helps in avoiding the keys to be stored in the proxy. Since, keys are derived dynamically when user logged-in, adversary

cannot derive keys of compromised system user’s keys that are not logged-in. This avoids any attacks based on available cipher text to adversaries.

• *Empirical results of PBKDF*

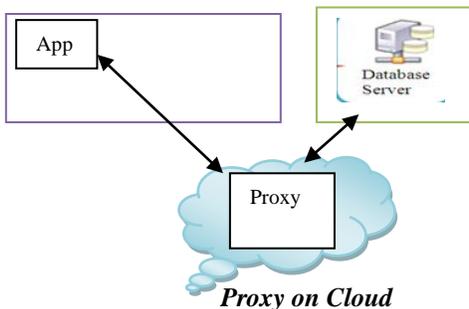
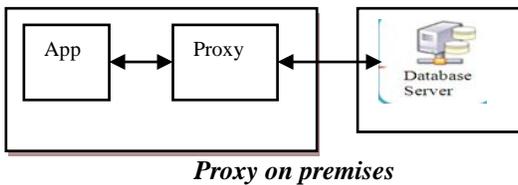
The empirical results of PBKDF are shown in the following table with different key length, number of rounds and SHA techniques. The key generation process is one time only when user login into the application therefore the timings for generation of keys with PBKDF are reasonable as shown in the table 1

| Sno | Password(Bytes) | Length of salt(bits) | Key Length(bits) | Number of rounds | SHA    | Timings in sec |
|-----|-----------------|----------------------|------------------|------------------|--------|----------------|
| 1   | 8               | 128                  | 128              | 1000             | sha1   | 0.010000       |
| 2   | 8               | 128                  | 128              | 100000           | sha1   | 0.900000       |
| 3   | 8               | 128                  | 256              | 1000             | sha1   | 0.020000       |
| 4   | 8               | 128                  | 256              | 100000           | sha1   | 1.680000       |
| 5   | 8               | 128                  | 128              | 1000             | sha512 | 0.010000       |
| 6   | 8               | 128                  | 128              | 100000           | sha512 | 0.520000       |
| 7   | 8               | 128                  | 256              | 1000             | sha512 | 0.010000       |
| 8   | 8               | 128                  | 256              | 100000           | sha512 | 1.040000       |

Table1: Performance of PBKDF

• *Proxy deployment*

The proxy can be deployed within the premises of the organization and also in third party service provider’s environment (Cloud). When deployed on the cloud proxy should be managed by the organizational administrator.



**4. Proposed Procedure for Key Generation and usage in Proxy based security solutions**

To decide the key that should be used for each data item, developers annotate the application’s database schema to

the proxy. The proxy maintains database schema and their access rights to each user i.e. what data item can be accessed by each user.

1. On user login into the application using password, application passes username and password to proxy in a SQL statement and the proxy generate the required encryption keys (MK) using PBKDF using this password.
2. The generated encryption key (MK) used for encryption and decryption of the sensitive data or used for deriving further encryption and decryption keys at different levels( databases, tables, columns) for that login user in the Proxy.
3. While encrypting the sensitive data using MK, PRIV\_TO and OPEN\_TO relations are maintained in the proxy.
4. The PRIV\_TO: Denotes that if the data item (D) is encrypted with user X key then data item D is PRIV\_TO user X. The OPEN\_TO: Denotes that data if data item D is also accessible to the Y and Z the users X key used for encrypting the data item D is encrypted with public keys of Y and Z and stored separately.
5. When data need to be accessed as per OPEN\_TO relation, use the password of the login user and generate the required private key (PrK) using PBKDF.
6. Decrypt the owner of the shared data encryption key with generated private key (PrK) and use that key for accessing the shared data.

Public-key for each user is generated and stored in the proxy using PBKDF2. No public key corresponding private key will be stored in the proxy and it will be generated in the proxy when user logins. The private key generation procedure explained below.

**Procedure for generating password based asymmetric key**

1. Generate a symmetric key ‘MK’ using PBKDF2 for given password.
2. Use this key ‘MK’ as seed for Pseudo Random Number Generator to produce a random number (RN). PRNG is deterministic (same seed implies same output sequence) therefore produces same random number (RN) for the given seed ‘MK’.
3. Generate a key pair using this Random number (RN) with an algorithm used to generate the asymmetric key like Elgamal or any other public-key algorithms.
4. The algorithm generates a same pair of keys(public and private) for a given source password, whenever application user login along with generation of symmetric key, a private key can be generated.

| Sno | Random number size(bits) using openssl rand() function | Timings in sec |
|-----|--|----------------|
|     |  |                |

|   |      |                     |                                      |
|---|------|---------------------|--------------------------------------|
| 1 | 128  | real<br>user<br>sys | 0m0.003s<br>0m0.001s<br>0m0.001s     |
| 2 | 256  | real<br>user<br>sys | 0m0.004s<br>0m0.001s<br>0m0.003s     |
| 3 | 512  | real<br>user<br>sys | 0m0.008s<br>0m0.001s<br>0m0.002s     |
| 4 | 1024 | real<br>user<br>sys | 0m0.008s<br>0m0.004s<br>0m0.004s     |
| 5 | 2048 | real<br>sys         | 0m0.008suser<br>0m0.006s<br>0m0.002s |
| 6 | 4096 | real<br>sys         | 0m0.008suser<br>0m0.006s<br>0m0.002s |

Table 2: Random number generator timings

Advantage of this technique is that requirement of storing of keys will be avoided at the proxy level. Hence, all possible attacks on stored encrypted keys at proxy level become invalid.

In CryptDB, key derivation or decryptions are based on the password provided by the application to the proxy. Suppose, a malicious Cross Site Script is present in the application server hosted. When authentic user is trying to run web application in his browser accessed from the web server. The malicious Cross Site script will execute in the authentic user browser, which may be intended to steal the user name and passwords of the application users. If these passwords are compromised by using XSS then user becomes a legitimate user and accesses all the confidential data, because user application passwords are associated with key decryption or derivation. To avoid this situation, along with the user name and password an OTP need to present to the user after authentication, a pop-up window will be displayed to the user from the Proxy to receive the OTP and this should be used for derivation of the encryption/decryption keys. Even if the username and passwords are compromised at the user level (Application), deriving the encryption/decryption keys becomes very difficult, because OTP is presented to the user dynamically from the proxy server and without OTP adversary cannot able to derive the right key. Therefore, XSS attacks in the context of the proxy type (CryptDB) of data protection mechanism, OTP in combination with password handle effectively.

- **Change of password:**

Whenever a password is changed, any data that is protected by the retiring password shall be recovered (e.g., decrypted) using the appropriate DPK that is associated with the retiring password, and then re-protected (e.g., encrypted) using the appropriate DPK that is associated with the revised password. This activity may be done online or offline to the proxy.

- **Password forgets**

It is very likely that passwords may be forgotten which are used for generation of encryption keys. Since, generation of right encryption key is depends on right password, therefore, the parameters(password, salt) which are used for generation of keys have to be securely stored on other than proxy system.

- **Integrity of query and results**

CryptDB is not checking the integrity of the query and results received from user application and the database server respectively. To identify on transition modifications of the results, before sending to the proxy the data hash has to be pre –calculated and attach to the data as an integrity header. On receiving of query or data, proxy computes hash of data and compares with received hash of data. This way proxy can ensure the integrity of query and results. Even further to ensure authentication and confidentiality of results, use of CCM mode of block cipher which provides authentication, confidentiality of user query and result of query in the proxy on the top of prevailing security.

- **Adding new application users to proxy**

Adding a new application user, the proxy needs to know credentials of that user and OPEN\_TO relation with other users for the shared data. To add new application user, application sends new user password and OPEN\_TO relation information with other user data to the proxy.

- **Possible attacks on cipher text in the proxy for stored keys[14]**

- i. Ciphertext-only attack: In this attack the attacker knows only the ciphertext to be decoded. The attacker will try to find the key or decrypt one or more pieces of cipher text (only relatively weak algorithms fail to withstand a ciphertext-only attack).
- ii. Known plaintext attack: The attacker has a collection of plaintext-ciphertext pairs and is trying to find the key or to decrypt some other ciphertext that has been encrypted with the same key.
- iii. Chosen Plaintext attack: This is a known plaintext attack in which the attacker can choose the plaintext to be encrypted and read the corresponding ciphertext.
- iv. Chosen Ciphertext attack: The attacker has the able to select any ciphertext and study the plaintext produced by decrypting them.
- v. Chosen text attack: The attacker has the abilities required in the previous two attacks.

All the mentioned attacks are trying to capture plain text corresponding to cipher text in different ways. If encrypted keys are stored in the proxy under the compromised proxy system all these attacks may be applicable and tries to gain the access to un-logged users keys stored in the proxy. Our proposed solution will mitigate all these attacks on the ciphers text.

## 5 Conclusions

Usage and protection of third party database service has become a critical requirement for many organizations. Organizational sensitive data passes through three stages in the database such as data at rest, transition and data in usage. Data in first two stages are adequately protected and to maintain complete confidentiality of organizational sensitive data, data in third state need to be protected. Data in usage i.e. data in computation may be protected by directly performing operations on encrypted data without decrypting. Sensitive data need to be encrypted before hosting onto third party database and perform required OLTP and OLAP operations on encrypted data in the database server based on the application SQL query and encrypted results will be send back to the application/proxy and where application/proxy decrypts the results. In this paper, we have presented a key management technique for proxy based security solutions where required keys are generated in the proxy using mechanism based on NIST recommended SP800-132 key management. These keys used to encrypt the data using encryption algorithms in the proxy such way that direct operations on encrypted data are possible in the database server. This algorithm generates required encryption keys used in the proxy from user provided password from application dynamically. This eliminates storing the encryption keys in the proxy hence avoids all kinds of attacks on the stored encryption keys in the proxy. Unlike CryptDB, no encryption keys are stored in the proxy and required encryption keys are generated in the proxy when user logins into the application and all the encryption keys will be deleted in the proxy once user logout from the application. This kind of security solution satisfies much regulatory compliance by not storing the keys.

## 6 References

- [1] Raluca Ada Popa, Catherine M. S. Redfield, Nikolai Zeldovich, and Hari Balakrishnan, "CryptDB: Protecting Confidentiality with Encrypted Query Processing", SOSP '11, October 23–26, 2011, Cascais, Portugal.
- [2] NIST Special Publication 800-132: Recommendation for Password-Based Key Derivation Part 1: Storage Applications by Meltem Sönmez Turan, Elaine Barker, William Burr, and Lily Chen.
- [3]<http://www.verizonenterprise.com/pcireport/2014/>
- [4]<http://web.townsendsecurity.com/encryption-key-management-resources/>
- [5]<http://townsendsecurity.com/products/encryption-key-management>
- [6] Craig Gentry. Fully homomorphic encryption using ideal lattices. In STOC, pages 169-178
- [7] "Building web applications on top of encrypted data using Mylar", Raluca Ada Popa, Emily Stark, Jonas Helfer, Steven

Valdez, Nikolai Zeldovich, M. Frans Kaashoek, and Hari Balakrishnan MIT CSAIL and †Meteor Development Group.

[8] SPORC: Group Collaboration using Untrusted Cloud Resources Ariel J. Feldman, William P. Zeller, Michael J. Freedman, and Edward W. Felten Princeton University

[9] Depot: Cloud storage with minimal trust Prince Mahajan, Srinath Setty, Sangmin Lee, Allen Clement, Lorenzo Alvisi, Mike Dahlin, and Michael Walfish The University of Texas at Austin, [fuss@cs.utexas.edu](mailto:fuss@cs.utexas.edu)

[10] Secure Untrusted Data Repository (SUNDR) Jinyuan Li, Maxwell Krohn\*, David Mazieres, and Dennis Shasha` NYU Department of Computer Science

[11] An Oracle White Paper July 2012 Oracle Advanced Security Transparent Data Encryption Best Practices

[12]<http://www.ciphercloud.com/technologies/encryption/>

[13] <https://securosis.com/tag/navajo+systems>

[14][http://www.facweb.iitkgp.ernet.in/~sourav/Attacks\\_on\\_cryptosystems.pdf](http://www.facweb.iitkgp.ernet.in/~sourav/Attacks_on_cryptosystems.pdf)



**Kurra Mallaiah** has been working in Defence Research and Development Organization (DRDO) last 13 years as a Scientist. Presently, he is with Advanced Numerical Research and Analysis Group (ANURAG) lab at Hyderabad as a Scientist 'D' and pursuing Phd in the field of data security in Osmania University under the supervision of Prof. Ramachandram, principal, University college of engineering, Osmania university. Prior to joining ANURAG, he was working in Indian Naval Ship (INS), Shivaji, Indian Navy as a Scientist 'C'. He served/acting as a reviewer for international journals. He has published papers in international journals and Conferences in the field of data security. Kurra Mallaiah holds a B.Tech and M.Tech degrees in Computer Science and Engineering. He is a Member of ACM, CSI, Defence Science Journal, Internet Society, and IAENG.



Dr. S. Ramachandram (1959) received his bachelor's degree in Electronics and Communication (1983), Masters in Computer Science (1985) and a Ph.D. in Computer Science (2005). He is presently working as a Professor and Principal in University College of Engineering, Osmania University, and Hyderabad, India. His research areas include Mobile Computing, Grid Computing; Cloud computing, Server Virtualization and Software Engineering. He has authored several books on Software Engineering, handled several national & international projects and published several research papers at international and national level. He also held several positions in the university as a Chairman Board of Studies, Nodal officer for World Bank Projects and chair of Tutorials Committee. He is a member of Institute of Electrical and Electronic Engineers (IEEE), Computer Society of India (CSI) and Institute of Electronics and Telecommunication Engineers (IETE).

# Time of flight Measurement Method to Determine the Milk Coagulation Cut Time

Mourad Derra, Abdellah Amghar, and Hassan Sahсах

*Abstract*—The coagulation time is often used as a reference to determine the time of cutting the gel, in order to expel whey trapped in the pores of the gel, [1]. Cutting time means that the gel has reached a certain firmness allowing passage from the enzymatic phase to the physico-chemical phase of the cheese making process. Therefore, cutting the gel earlier or later will have negative effects on the quality of the final product. So, optimal evaluation of the coagulation time is necessary to maximize qualitative and quantitative cheese yields. In this work a non-destructive ultrasonic technique is developed to monitor in real time the coagulation process of renneted milk in order to determine the coagulation time. The latter is determined with high precision by exploiting changes of the time of flight in the coagulating milk. We have developed a non-invasive technique that uses a single transducer, what is very important in the food industry.

*Keywords*—Real time control, time of flight, milk coagulation, coagulation time.

## I. INTRODUCTION

**T**HE conversion of milk into cheese involves four steps: coagulation, drainage, salting and ripening. The coagulation step is the most important in the process of cheese making, it consists of two consecutive phases: enzymatic phase and physicochemical phase. This step is generating an increasing interest in the cheese industry since it plays a decisive role in determining the quality of the final product. Therefore, mastery and good control of this stage of cheese production remains the major concern of both researchers and industrialists. In the manufacture of the majority of cheese varieties, milk proteins are coagulated to convert milk from liquid to semi-solid state (gel) in which the fat globules, the water and the materials are trapped in pores. When a certain strength of the gel was achieved due to the coagulation process, the gel was cut into slices of about 7 mm cubic. The coagulation matrix shrinks and causes the expulsion of whey from cubes (syneresis), thus leading to a two-phase system: curd and whey. After that the separated curd from whey is processed more to achieve a cheese product. Current practice in the cheese industry is to cut the gel after a fixed time of the enzymatic reaction or to rely on the subjective judgment of an operator to determine the correct cutting time. Conventionally, many manufacturers of cheese cut the gel 30 minutes after adding rennet to the milk in order to meet the factory programs, [2]. This practice is uncertain because many factors affecting the gel properties are not constant. For this reason, a more objective determination of cutting time will improve the cheese making and maximize

performance, [3]. Several studies aimed at controlling in real time the process of milk coagulation were carried out to determine the clotting time and therefore the perfect time for slicing the gel. These works are based on electrical, [4], thermal, ([5]-[7]), optical, ([8]-[11]), and viscometric methods, ([12]-[15]). However, most of these methods have a destructive character that their direct contact with the coagulum causes the deformation of the gel what limits the quality of the final product. To find more effective ways having a non-destructive nature, then the use of ultrasound proves interesting, ([16]-[21]). Because of their ease of placing in-situ and their non destructive properties, the techniques of ultrasonic wave propagation seem particularly well suited to monitor the gelation of the milk.

Ultrasonic velocity and attenuation have become very valuable tools for studying the physical properties of matter. For this reason many studies on milk coagulation, ([16]-[21]) are based on the monitoring of these two quantities (velocity and attenuation) during the clotting process using the technique of transmission. However, these studies use only the attenuation evolution to determine the clotting time. Using an ultrasonic resonator technique Buckun and al. [22] measured with a good resolution the velocity of phase and attenuation during coagulation. However, this technique has not been used to determine the clotting time. Another method of F.Bakkali and al. [23] use an ultrasonic pulse technique by reflection for determining the clotting time from the phase velocity changing by calculating the second derivative of the polynomial function obtained by smoothing the experimental curve of the phase velocity [23].

In this work, we have developed a non-destructive ultrasonic technique for monitoring changes in the time of flight during the enzymatic coagulation of milk.

## II. MATERIALS AND METHODS

### A. Materials

We conducted our experiments with skim milk powder from the same package, purchased commercially. For 90 g of water we added 13 g of skimmed milk powder. Everything is highly homogenized and placed in a thermostatted water bath at the desired temperature for 30 minutes before being rennet. The rennet used is sold in pharmacies under the name Caille-lait universel 0.22 g / l , a product of society COOPER Morocco. Rennet tablet is dissolved separately in a little water. The proportions recommended by the manual rennet have been respected; the equivalent of one tablet per liter of reconstituted milk. Homogenization of the whole milk + rennet

M. Derra, A.Amghar and H.Sahсах are with Laboratory of Metrology and Information Treatment, Faculty of Science, Agadir, Morocco. email: mourad.derra7@gmail.com

TABLE I  
 EXPERIMENTAL CONDITIONS.

| Conditions (notation)                              | Temperature (° C) | Rennet concentration (g/liter milk) |
|--|-------------------|-------------------------------------|
| Standard temperature (ST) and standard rennet (SR) | 37                | 0.22                                |
| Standard temperature (ST) and low rennet (LR)      | 37                | 0.20                                |
| Standard temperature (ST) and high rennet (HR)     | 37                | 0.24                                |
| High temperature (HT) and standard rennet (SR)     | 42                | 0.22                                |
| Low temperature (LT) and standard rennet (SR)      | 32                | 0.22                                |

is necessary to avoid creating bubbles. The acquisition of the signal received by the transducer at every step of one minute is launched on PC. The data is then processed in the LabVIEW program in order to determinate the viscoelastic parameters of milk. As the temperature and the concentration of rennet are the main factors affecting the enzymatic phase of coagulation of milk, ([17]-[24]) three levels of each of these parameters were considered during the experiments. They represent a standard level and two levels, one is below and the other is above (see Table 1).

### B. Experimental device

The figure 1 shows the device of measurement used. A center frequency transducer of 5 MHz (0.5 in, crystal diameter, A309S-SU Model, Panametrics, Olympus) immersed in water, is used to make the ultrasonic impulse crosses the container enclosing the milk sample. The whole is in a thermostat tank. The sensor is connected to a generator pulse (Sofranel Model 5073PR, Sofranel Instruments) which plays the role of transmitter/receiver that sends the electric signal. The received signal after interfaces reflections : water / plexiglas, Plexiglas / milk and milk / glass, is amplified and digitized by a PicoScope. The different treatments applied to digital signal obtained are performed using a LabVIEW program, developed in this work and implemented on a personal computer.

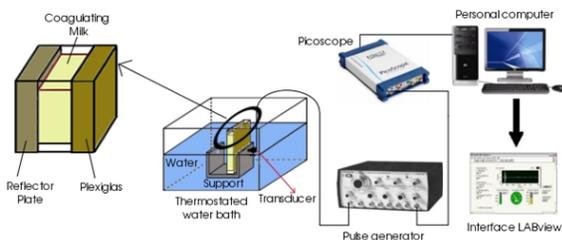


Fig. 1. Experiment setup

The incident signal follows the path shown by the diagram of figure 2.

### C. Control in real time under labVIEW interface of coagulating milk

We start the control by acquiring the formed signals on the computer based on an application programmed with labVIEW

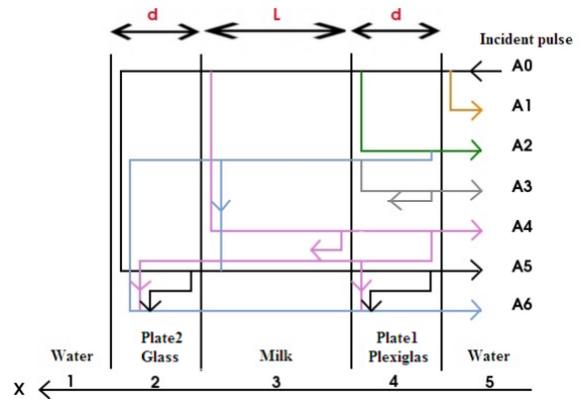


Fig. 2. Different paths of propagation

language. The realized program allows us to specify the number of wanted acquisitions during the experiment and the time between each other. This program captures each time 50 signals and realizes an averaging to neutralize noised signals. The resulting signal represents for the user a single acquisition (Figure 3).

This signal comprises a first part composed of three echoes:

- echo A1, linked to the specular reflection of the incident beam at the interface between water and the first face of the plate 1,
- the second echo A2 corresponding to the reflection at the interface between the second side of the plate 1 and the enclosed milk,
- the third echo A3 corresponding to the second back and forth in the plate 1.

The second part includes also three echoes, having traversed all the trapped milk in back and forth:

- the first echo A4 corresponds to reflection at the interface between the milk and the plate 2,
- the echo A5 represents a superposition of two echoes : one make a back and forth in the plate 1, and the other one in the plate 2,
- the third echo of the series A6 corresponds to a superposition of three echoes : an echo which made two back and forth to the plate 2, an echo that made two back and forth to the plate 1, and an echo corresponding to a back and forth in the plate 1 followed by a back and forth in the plate 2.

### D. Technique of measuring the time of flight

The time of flight is written as in the following formula :

$$t_v = \frac{1}{2} [(t_{4a} - t_{2a}) + (t_{4b} - t_{2b})] \quad (1)$$

## III. RESULTS AND DISCUSSIONS

Ultrasonic techniques developed until today to determine the clotting time, are based on the evolution of the phase velocity.

The latter, as it was used by F.Bakkali and al.[23] consist to use the second derivative of the phase velocity and consider the

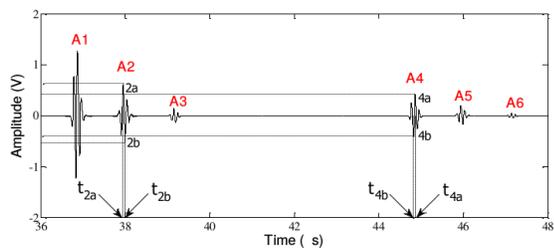


Fig. 3. Typical waveform of the reflected ultrasonic signal

extreme of this second derivative as the transition point which marks the end of the enzymatic phase of coagulation. This transition point has been identified as clotting time ([18],[19]).

In this work, we applied this method for determining the clotting time in our experimental results, and we compared it to our method based on exploitation of the the time of flight. The figure 4 present, in our standard experimental conditions, the change as a function time of the smoothed time of flight and its second derivative. The values of the clotting time measured from the second derivative of phase velocity are grouped for the different experimental conditions in table 2. The time of flight obtained at different rennet concentration and temperature is shown in figures (5 ; 6 ; 7 ; 8 ; 9).

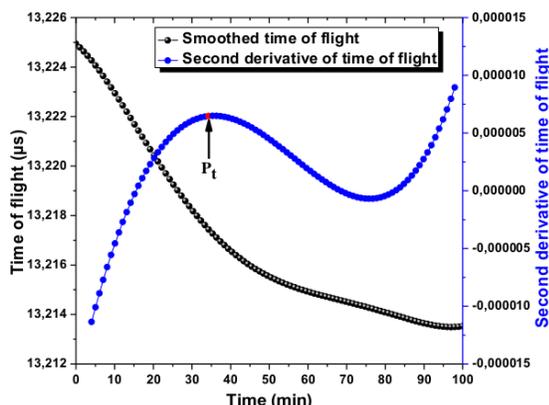


Fig. 4. Identification of the transition point from the second derivative of the time of flight under standard conditions

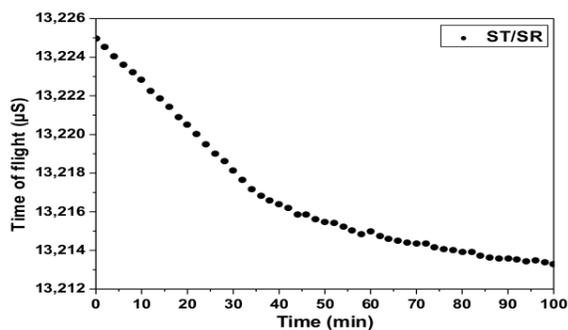


Fig. 5. Time of flight through milk coagulation at ST/SR

The method for determining the clotting time based on the

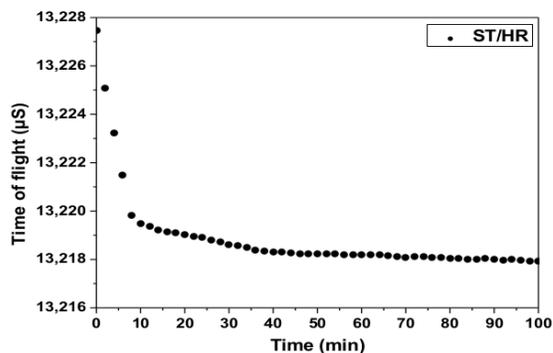


Fig. 6. Time of flight through milk coagulation at ST/HR

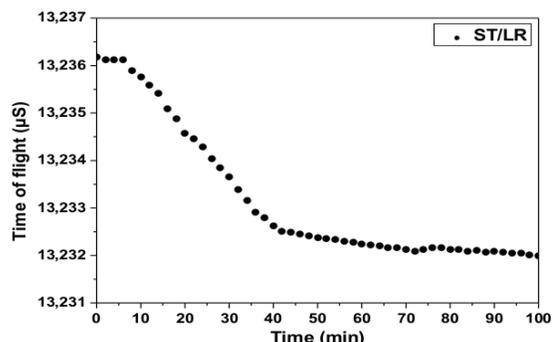


Fig. 7. Time of flight through milk coagulation at ST/LR

evolution of the time of flight remains one of the novelties of this work. The values of clotting time measured from the time of flight are grouped for different experimental conditions in table 2.

Note that the transition point at an elevated temperature (42 C) is earlier than in the case of the standard temperature (37 C). The same applies to a high rennet concentration, the transition point is earlier than in the case of a standard concentration. This indicates that the coagulation is much faster than the temperature or the rennet concentration are high.

#### IV. CONCLUSION

In the previous method the calculation of phases used to compute the phase velocity, two spectra whose phase varies between  $-\frac{\pi}{2}$  and  $\frac{\pi}{2}$  are obtained, however, it is necessary to achieve continuous phase spectra for the calculation of the speed of phase. Which required unwrapping the phases into

TABLE II  
CLOTTING TIME MEASURED FOR THE DIFFERENT EXPERIMENTAL CONDITIONS OF TEMPERATURE AND RENNET

| Condition | Phase velocity | Time of flight |
|-----------|----------------|----------------|
| ST/SR     | 35.40 0.93     | 34.95 1.04     |
| ST/HR     | 24.70 1.04     | 24.26 1.02     |
| ST/LR     | 36.38 0.61     | 35.86 0.45     |
| HT/SR     | 27.45 0.99     | 27.51 1.02     |
| LT/SR     | 39.81 2.40     | 39.38 2.83     |

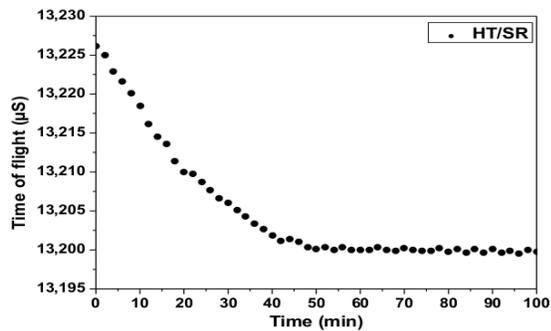


Fig. 8. Time of flight through milk coagulation at HT/SR

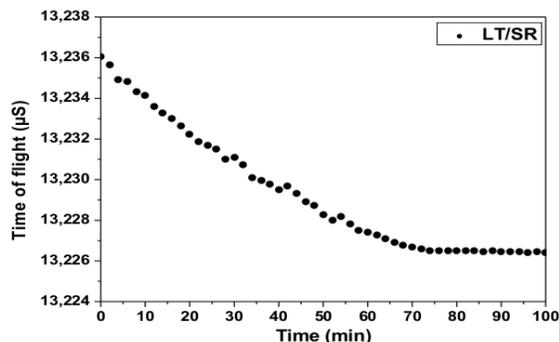


Fig. 9. Time of flight through milk coagulation at LT/SR

continuous form. This unwrapping needs a signal that contains a huge number of values. The new ultrasound technique developed in this work allows us to determine the clotting time, based on time of flight changes. Due to the simplicity of its conditions, this technique may replace the method of determining the coagulation time based on the evolution of the phase velocity.

So this technique proves to be well suited for this kind of control: it is non-destructive and non invasive, especially since it makes - using a program - a real-time control.

#### ACKNOWLEDGMENT

This work was supported by the Centre National pour la Recherche Scientifique et Technique (CNRST), funded by the Moroccan government.

#### REFERENCES

- [1] Weber, F. 1987. Curd drainage Cheesemaking. Science and Technology ed A Eck (New York: Lavoisier): ch 2.
- [2] Linklater, P. M., Porc. Symp., Kensington, Australia. 1966. Univ. New South Wales, Kensington, N. S. W. Australia.
- [3] Brynum, D.G. & Olson, N.F. J. 1982. Dairy Sci (65) : 2281.
- [4] Dejmeck, P. 1988. Precision conductometry in milk renneting. J. Dairy Res (56) : 6978.
- [5] Bruno, C., Grasso, G., Spagna, S., Matteo, M. & Micione, B. A. 1986. Comparison between dynamometric and thermal conductivity parameters: the milk clotting time. Sci. Tec. Casearia (37):457472.
- [6] Hori, T., Miyawaki, O. & Yano, T. 1989. In line measurements of milk clotting time by hot-wire method Lecture at ICEF5 (Cologne).
- [7] Passos, E. F., Monteiro, P. S., Oliveira, R. C., Martins, J. G. O., Alves, H. G. & Brandao, S. C. C. 1999. Predicting the cutting time of coagulating milk for cheese production using a heated thermistor. J. Food Sci (64): 879882.

- [8] Famelart, M. H. & Maubois, J. L. 1988. Comparison of refractive index and viscosity evolution during lactic gelification of milk Lait (68) : 112.
- [9] Payne, F. A. 1995. Automatic control of coagulum cutting time in cheese manufacturing. Appl. Eng. Agric (11): 691697.
- [10] Castillo, M., Payne, F. A., Hicks, C. L. & Lopez, M. B. 2000. Predicting cutting time of coagulation goats milk using diffuse reflectance: effect of pH, temperature and enzyme concentration. Int. Dairy J (10): 551562.
- [11] Herbert, S., Riaublanc, B., Bouchet, B., Gallant, D. J. & Dufour, E. 1999. Fluorescence spectroscopy.
- [12] Kopelman, I. G. & Cogan, U. 1976. Determination of clotting power of milk clotting enzymes. J. Dairy Sci (59): 196199.
- [13] Korolczuk, J. & Maubois, J. L. 1987. Computerised viscometric method for studying rennet coagulation of milk. J. Texture Studies (18): 157172.
- [14] Lopez, M. B., Jordan, M. J., Granados, M. V., Fernandez, J. C., Castillo, M. & Laencina, J. 1999. Viscosity changes during rennet coagulation of Murciano-Granadina goat milk Int. J. Dairy Technol (52): 102106.
- [15] Shulz, D., Seng, B. & Krenkel, K. 1999. Investigations into the combined enzymatic and lactic acid milk coagulation Milchwissenschaft (54): 363367.
- [16] Benguigui, L., Emery, J., Durand, D. & Busnel, J. P. 1994. Ultrasonic study of milk clotting Lait (74): 197206.
- [17] Ay, C. & Gunasekaran, S. 1994. attenuation measurements for estimating milk coagulation time. Trans. Am. Soc. Agric. Eng (37) : 857862.
- [18] Gunasekaran, S. & Ay, Chyung. 1996. Milk coagulation cut time determination using ultrasonics. J. Food Process.Eng (19) : 6373.
- [19] Gunasekaran, S. & Ay, Chyung. 1999. Evaluating milk coagulation with ultrasonics. Semin. Food Anal (4) : 161173.
- [20] Taifi, N., Bakkali, F., Faiz, B., Moudden, A., Maze, G. & Dcultot, D. 2006. Characterization of the syneresis and the firmness of the milk gel using an ultrasonic technique. Meas. Sci. Technol (17): 281287.
- [21] Izbaim, D., Faiz, B., Moudden, A., Malainine, M. & Aboudaoud, I. 2012. Ultrasonic Characterization of yogurt fermentation process. Proceedings of the acoustics Nantes Conference.
- [22] Buckin, V. & Cormor, S. 1999. High resolution ultrasonic resonator measurements for analysis of liquids. Semin. Food Anal (4) : 113130.
- [23] Bakkali, F., Moudden, A., Faiz, B., Amghar, A., Maze, G., Montero, F. & Akhnak, M. 2001. Ultrasonic measurement of milk coagulation time. Meas. Sci. Technol(12): 21542159.
- [24] Brule, G. & Lenoir, J. 1987. The coagulation of milk Cheesemaking. Science and Technology ed A Eck (New York : Lavoisier): ch 1.

# Ensurance and simulation of electromagnetic compatibility: recent results in TUSUR University

Talgat Gazizov, Alexander Melkozerov, Alexander Zabolotsky, Sergey Kuksenko, Pavel Orlov, Vasilii Salov, Roman Akhunov, Ilya Kalimulin, Roman Surovtsev, Maxim Komnatnov, Alexander Gazizov

**Abstract**—The paper presents and summarizes theoretical and practical results of recent electromagnetic compatibility (EMC) projects which are focused on the development of approaches, technologies and software for EMC of electronic equipment. Some representative results of EMC simulation and ensurance are presented. Outlook for applications of the obtained results in future projects of TUSUR University is given.

**Index Terms**—EM software, EMC, signal integrity, simulation

## I. INTRODUCTION

In the design of critical equipment it is necessary to take into account the requirements of the electromagnetic compatibility (EMC). These requirements continuously increase as the packaging density and upper frequencies of desired and noise signals in critical electronics grow. Actual EMC testing of the electronics and repetitive redesign due to failure to comply with increasingly strict EMC requirements considerably raise the price and duration of the design process.

A representative example of the critical equipment is a space vehicle. The EMC problem is especially of current concern for prospective space vehicles. Their distinctive features such as unified electronic modules based on the “system-on-a-chip” technology, unpressurised chassis and increased lifetime (up to 15 years and more) make the EMC ensurance and simulation even more difficult. In particular, increasing packaging density and, consequently, crosstalk in printed traces requires a signal integrity analysis. Unpressurised chassis worsens the shielding efficiency of the whole space vehicle for certain frequency ranges and requires special approaches to the simulation of chassis elements’ shielding. Increased to 15 years lifetime makes the provision of significantly overestimated interference immunity margin necessary since noise electromagnetic excitations may grow during this period to such high levels which can be hardly predicted.

Therefore, careful analysis of a wide range of signal integrity, power integrity and EMC issues must be performed. However, focus on the analysis problems without computer-aided synthesis and optimization often makes the design ineffective and leaves hidden the resources for its

improvement. To investigate these problems and develop appropriate solutions, a number of EMC projects was and being conducted at Tomsk State University of Control Systems and Radioelectronics (TUSUR University) for 2009–2016 years. Brief summary of new results on EMC simulation for space projects of TUSUR University, showing state of the art in the field to interested researchers has been presented recently [1]. The aim of this review paper is to present and summarize the results of the last projects and provide an outlook for applications of the obtained results in current projects. This paper is considerably extended: by representative data for the results only shortly described previously; by results not of simulation only but its application to ensure the EMC; by new results obtained not only in the space but other EMC projects.

## II. THEORETICAL RESULTS

### A. Quasi-static and Electromagnetic Analysis

Several derived analytic models [2] for time-domain response calculation of cascaded transmission line sections with capacitive loads have been implemented in the TALGAT software. Moreover, 4 new models for 2D configurations [3] and 2 new models for 3D ones [4] have been implemented for the analytical calculation of linear system matrix entries when obtaining the capacitive matrix of arbitrary structures. The program implementation allows computing the  $\mathbf{L}$ ,  $\mathbf{C}$ , and  $\mathbf{G}$  matrices for arbitrary interconnections of high complexity and density which are characteristic for the components of the “system-on-a-chip” type.

Simple, but representative example of the new 2D models usage is shown here on characteristic impedance ( $Z$ ) estimations for three cases of a differential pair (Fig. 1). For ordinary case of rectangular conductors (Fig. 1(a))  $Z=50.42 \Omega$ . For the more real case of the curvilinear upper corners (Fig. 1(b))  $Z=51.42 \Omega$  that is by 2% more. At last, for previous case covered by CARAPACE EMP110 solder mask of thickness  $20 \mu\text{m}$  and polyparaxylylene water resistant layer of thickness  $15 \mu\text{m}$ , as taken from real world printed circuit board (PCB) of spacecraft, (Fig. 1(c))  $Z=48.53 \Omega$  that is by 4% less than in the first case. Such deviations are considerable for signal integrity, but can be controlled using the new models.

At last, the classical algorithm [5] using RWG-functions for electromagnetic analysis of arbitrary structures consisting of conductive patches has been improved by means of the calculation of the integrals using analytic formulas [6]. Comparative results of the first entry of  $\mathbf{Z}$  matrix computation by these formulas and Newton–Kotes integration for a plate of width  $0.5\lambda$  excited by plane wave are shown in Table I.

---

Review of theoretical results were supported by RFBR grant 14-29-0925, practical results were supported by the state contract 8.1802.2014/K of the Russian Ministry of Education and Science, algorithms for accelerated linear systems solution were developed under RFBR grant 14-07-31267, analytical review of modal technologies is carried out at the expense of RSF grant 14-19-01232 in TUSUR.

All authors are with Department of Television and Control of Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia.

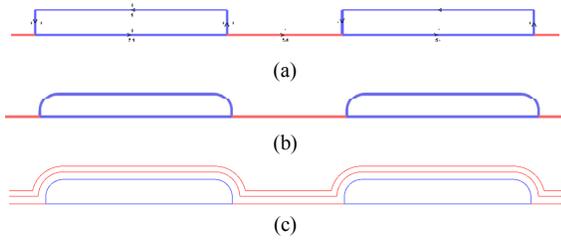


Fig. 1. Scaled cross sections of differential pairs of (a) rectangular, (b) curved, (c) curved and covered conductors

TABLE I.  
THE RESULTS OF COMPUTING THE REAL AND IMAGINARY PARTS OF  $Z_{11}$

| Computing the integrals   | Re( $Z_{11}$ ), $\Omega$ | Im( $Z_{11}$ ), $\Omega$ | $t$ , ms |
|---------------------------|--------------------------|--------------------------|----------|
| Numerical, accuracy 1     | $4.814 \cdot 10^{-8}$    | $-5.327 \cdot 10^{-9}$   | 95       |
| Numerical, accuracy 0.1   | $4.902 \cdot 10^{-8}$    | $-5.326 \cdot 10^{-9}$   | 453      |
| Numerical, accuracy 0.01  | $5.101 \cdot 10^{-8}$    | $-5.324 \cdot 10^{-9}$   | 1322     |
| Numerical, accuracy 0.001 | $5.293 \cdot 10^{-8}$    | $-5.327 \cdot 10^{-9}$   | 6000     |
| Analytic                  | $5.328 \cdot 10^{-8}$    | $-5.327 \cdot 10^{-9}$   | 502      |

From Table I one can observe that when increasing the integration accuracy the results obtained converge to those obtained by analytic formulas. However, the time of numerical integration increases (12 times for the integration accuracy equal to 0.001), whereas the time of computing the integrals by analytic formulas remains constant.

### B. Solution of Linear Algebraic Systems

The modeling in the range of parameters is often required. Example of the quasi-static problems is the capacitance matrix  $\mathbf{C}$  calculation by method of moments for the frequency dependent dielectric constant. In this case, for each frequency a linear algebraic system is solved (order of the matrix  $N$  is defined by a sum of subintervals on conductor  $N_C$  and dielectric  $N_D$  boundaries) with  $N_{COND}$  (number of conductors in the structure) right hand side vectors. Thus, the time of calculation is proportional to the number of frequency points. However, for this case (of the dielectric constant variation), only the elements in the lower part of the main diagonal of the matrix (corresponding to the dielectric subintervals) are filled again for each frequency point. Therefore, the exploiting the incomplete change of the linear system matrix will significantly reduce the time of multiple calculations.

For this aim the block LU-decomposition is useful, wherein the original matrix is divided into blocks:  $\mathbf{S}_{11}$  of size  $N_C \times N_C$ ,  $\mathbf{S}_{12} - N_C \times N_D$ ,  $\mathbf{S}_{21} - N_D \times N_C$ ,  $\mathbf{S}_{22} - N_D \times N_D$ . Only  $\mathbf{S}_{22}$  entries are recalculated for each frequency point. The algorithm of the block LU-decomposition:

1. Assign  $\mathbf{U}_{11} = \mathbf{S}_{11}$ ,  $\mathbf{U}_{12} = \mathbf{S}_{12}$ .
2. Calculate  $\mathbf{L}_{21} = \mathbf{S}_{21} \mathbf{U}_{11}^{-1}$ .
3. Calculate  $\mathbf{U}_{22} = \mathbf{S}_{22} - \mathbf{L}_{21} \mathbf{U}_{12}$ .

Obviously, the changing only the  $\mathbf{S}_{22}$  diagonal entries changes  $\mathbf{U}_{22}$  entries, because of (assigning  $\mathbf{U}_{11} = \mathbf{S}_{11}^{-1}$  at the first step) all the remaining blocks ( $\mathbf{U}_{11}$ ,  $\mathbf{U}_{12}$ ,  $\mathbf{L}_{21}$ ), including product  $\mathbf{L}_{21} \mathbf{U}_{12}$ , remain unchanged. Thus, for the  $M$  calculations the only time consuming expansion of linear system matrix and subsequent  $M-1$  computations of  $\mathbf{U}_{22}$  block are needed. Obtained expression for the maximum acceleration for the block LU-decomposition at  $M \rightarrow \infty$ :

$$\beta = \lim_{M \rightarrow \infty} \frac{M \cdot T_{LU}}{T_1 + (M-1) \cdot T_S} = \lim_{M \rightarrow \infty} \frac{M \cdot T_{LU}}{T_1 + M \cdot T_S - T_S} = \frac{T_{LU}}{T_S},$$

where  $T_1$  – time of the first solution, which includes inverting the  $\mathbf{U}_{11}$  of size  $N_C \times N_C$  and the subsequent solution of the linear system with finding right hand side vector;  $T_S$  – time of computing the  $\mathbf{U}_{22} = \mathbf{S}_{22} - \mathbf{L}_{21} \mathbf{U}_{12}$  and further solving the linear system. It is evident from the expression that the more the  $M$ , the smaller the acceleration depends on the time of the first solution. Moreover, the acceleration to a large extent depends on the size of the  $\mathbf{S}_{22}$  block. Thus, for large  $M$ ,  $N$  and small  $N_D$  the maximum acceleration of the multiple calculations can be obtained, whereas for small  $M$  and large  $N_D$  the acceleration will be negligible or not at all. The block LU-decomposition has been implemented and investigated for acceleration of multiple solution of linear equations with partly changing matrix. Acceleration by factor of more than 2 has been obtained for the simulation of the SNP 339 type connector. For particular structures, the acceleration can be up to factor of 35.

Obviously, if the analysis is required when changing the geometrical parameters of the structure, the matrix entries will vary in arbitrary places, and therefore the above-mentioned approach is not applicable. To overcome this problem a use of iterative methods was proposed.

The algorithm for multiple iterative solution of the linear system with partially changing matrix was presented in [7]. In this algorithm, the preconditioner matrix  $\mathbf{M}$  is formed from the first linear system. Further, this matrix (without recalculation) is used for solving the following linear systems, thereby reducing the total solution time with acceptable accuracy. Finally, it was supposed that a similar algorithm can be applied when changing the sizes of the structure being analyzed. As a first step in this direction the reduction of the residual norm was investigated for solving the 10 linear systems, obtained by small changes for several parameters of a structure [8]. Example of these calculations for dielectric height ( $h$ ) of microstrip line (Fig. 2) is presented in Fig. 3.

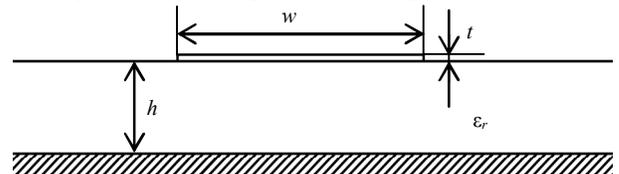


Fig. 2. Cross section of structure under consideration

All obtained results show that it is possible to use as a preconditioner the factorized matrix computed from initial parameters for multiple iterative solution of linear systems when changing the any parameters. But as the difference between the values of these parameters increases (by factor 2) the number of iterations also increases. Therefore, it is important to know the maximum number of iterations when the time required for solving by the iterative method is less than by the direct method. If the solution does not converged, then it is possible to recalculate preconditioner to get the solution convergence for total range of changing the parameters. In any case, the proposed approach may decrease considerably the time of multiple iterative solution of linear systems. Estimates of the speed up may be easily performed.

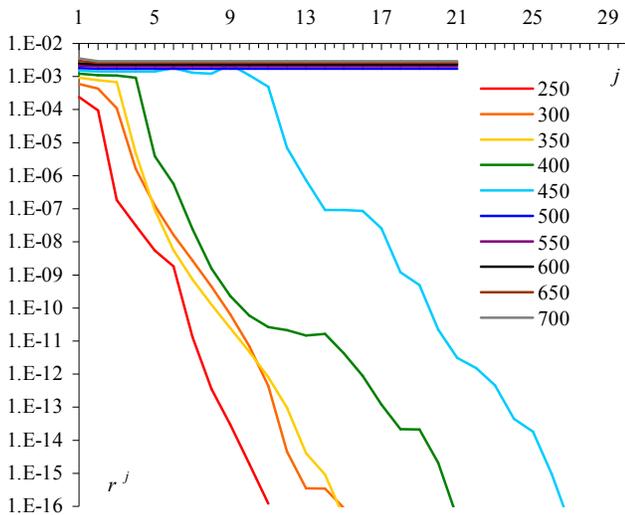


Fig. 3. Dependence of relative residual norm ( $r^j$ ) on the number of iterations ( $j$ ) for  $h=250, 300, \dots, 700 \mu\text{m}$

It was shown that the use of compressed sparse row format (CSR) for storing a preconditioner matrix is effective to reduce the computational cost. Acceleration of iterative solution of linear systems with dense matrices using sparse matrix storage formats has been considered in details [9]. Formulas for comparing the sparse matrix storage formats have been derived. An iterative algorithm for solving linear algebraic systems using sparse row format for storing prefiltered preconditioners has been designed. A modification of the sparse row format leading to 1.14–1.23-times speed-up for matrices of order 1000 has been suggested. It has been demonstrated that as opposed to the usual storage format, the sparse row format provides for 1.5–1.6-times speed-up in solving the linear systems of orders 4800, 6000, and 8000. The use of the obtained results allows one to reduce both memory and time requirements in solving large-scale problems with dense matrices.

Then, improvements to the ILU(0) factorization algorithm for preconditioning linear algebraic systems with dense matrices have been suggested [10]. (The preconditioner is stored in compressed sparse row format.) On the example of the problem of computing the electrical capacity of two stripes, it has been demonstrated that the modifications proposed provide for a significant reduction of the time for computing the ILU(0) preconditioner (up to 4 times) and for solving the preconditioned linear system (up to 2.5 times).

On real PCB structure problems, a new investigation has been performed in order to reveal the optimal value of the main parameter (drop tolerance) of the iterative solution of linear systems. The algorithm for calculation of capacitance matrices of structures of conductors and dielectrics using the method of moments has been improved for case of multiple calculations. The improved algorithm has been shown to work up to 4 times faster than the initial one.

The possibility of multiple iterative solution of linear systems was further investigated for computing the capacity of microstrip line in the wide ranges of its sizes.

To accelerate the iterative process two ways were considered. The first one is a use of a previous linear system solution (vector  $\mathbf{x}_{i-1}$ ) as an initial guess for a following solution (vector  $\mathbf{x}_i^0$ ), i.e.  $\mathbf{x}_i^0 = \mathbf{x}_{i-1}$  (for the first system a unit vector are used). The second one is the use of preconditioning matrix  $\mathbf{M}$ , obtained by solving the first linear algebraic equation, i.e.  $\mathbf{M}_i = \mathbf{M}_0$ . In computational experiments four options were used: in option 1 acceleration was not used. In options 2 and 3, these ways were used separately, and in option 4 these ways were used together.

The previous structure (Fig. 2) was investigated. The aim of the experiment was to evaluate the time expenses required for the calculation of 100 capacitive matrices obtained by changing one of the dimensions of the structure: dielectric height  $h$  (in the range of 12–112  $\mu\text{m}$  or 933%); conductor width  $w$  (in the range 18–118  $\mu\text{m}$ , or 656%); conductor height  $t$  (in the range of 6–106  $\mu\text{m}$ , or 1767%). The number of segments on each boundary of structure has not changed, which allows for constant order  $N=1600$  of the linear system matrices for correct comparison. As iterative method the BiCGStab method was chosen. Iterations were continued until the relative norm of the residual vector was more than  $10^{-8}$ . Gaussian elimination was used for comparison.

Ratios of the total (for 100 linear systems) solution times by Gauss elimination and by iterative method are shown in Table II for all options. Calculations have demonstrated the effectiveness of the proposed acceleration ways. The number of iterations when using the option 4 is minimal, that reduces the total time of 100 of linear systems solution by factor about 5–12 and proves the effectiveness of the combined usage of acceleration ways.

TABLE II.  
TOTAL ACCELERATION OF MULTIPLE SOLUTIONS  
OF 100 LINEAR SYSTEMS FOR ALL OPTIONS

| Changed parameter | Option 1 | Option 2 | Option 3 | Option 4 |
|-------------------|----------|----------|----------|----------|
| $h$               | 0.48     | 1.32     | 6.49     | 11.77    |
| $w$               | 0.31     | 1.15     | 5.87     | 10.98    |
| $t$               | 0.37     | 1.28     | 2.87     | 4.92     |

### C. Modal Filtration

Theoretical investigation of modal filtration in the printed circuit boards (PCBs), cables and separate modal block-filters led to the following results [11–13]. An important condition for the choice of the resistive loads at the ends of a modal filter (MF) section has been found which provides for equal magnitudes for the pulses decomposed at the filter output. An analytic expression for calculation of normalized magnitudes of the decomposed pulses as a function of wave impedance of even and odd modes has been derived. The dependency of these magnitudes on the coupling in the MF line has been demonstrated. Analysis of the power which is dissipated by resistors at the ends of active and passive conductors of multi-stage MF has been performed [14]. The analysis showed how one can select resistors by their dissipation power for given excitation parameters. A possibility to use widely available flat power cables as a protection against dangerous pulses by

means of decomposition into pulses with smaller magnitude has been revealed [15, 16]. This device has been shown to be radiation-resistant because it contains no semiconductor elements; cheap because besides the cable it contains just resistors; light and reliable since even short-circuits or gaps of cable conductors can be used instead of resistors. A modal block-filter structure with optimal parameters has been chosen. An experiment with a single-stage modal block-filter has been performed which validated the outcome of the theoretical investigation.

Cross section of the simplest (one-layer) planar structure comprising three conductors on one side of a dielectric substrate is shown in Fig. 4. Left conductor is active, central conductor is reference, and right conductor is passive.

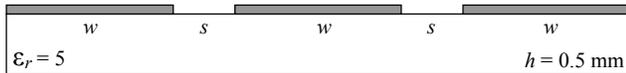


Fig. 4. Cross section of (one-layer) line section planar structure.

The typical length of printed coupled lines meandered on usual double-sided printed circuit board may be roughly assumed as 10 m for square of 1 dm<sup>2</sup>. Therefore the difference of per unit of length delays of 1.3 ns/m (multiplied by the length) for these lines will allow decomposition of the pulses shorter than 13 ns. Use of two-layer structures is more preferable because the higher value of the difference of per unit of length delays permits to decompose longer pulses or, alternatively, to shorten the minimum length of a cable for previous pulse durations. For example, the improvement may be almost by factor of 1.5 in comparison with the one-layer structure. Use of high permittivity dielectrics allows improving the presented results proportionally to  $\sqrt{\epsilon_r}$ , by factor of 6.

Cross sections of three conductor flat cables are shown in Fig. 5. All types of the cables have been classified in two kinds: with and without air gaps in cross section.

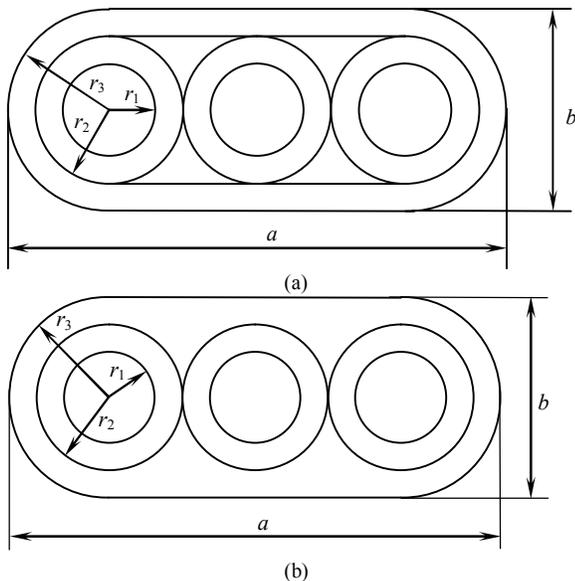


Fig. 5. Cross sections of cables with (a) and without (b) air gaps.

The typical length of low-voltage power cables may be assumed 10 m for domestic (room) and 100 m for floor (house)

applications. Then the difference of per unit of length delays 0.3 ns/m (multiplied by the length) for these cables will allow dividing the pulses shorter than 3 ns and 30 ns accordingly. Use of cables without air gaps is more preferable because the bigger value of the difference 0.5 ns/m permits to extend the pulses up to 5 ns and 50 ns accordingly or, alternatively, to shorten the minimum length of a cable for previous pulse durations.

Modal filter consists of  $N$  cascaded sections having the same cross sections but twice length (Fig. 6).

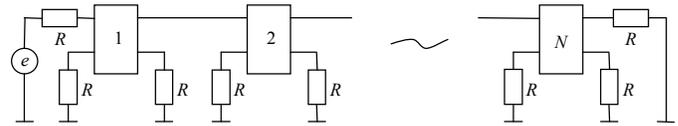


Fig. 6. Schematic of the modal filter.

To illustrate the functioning of the modal filter, we consider an example of the time-domain response calculation for excitation between signal (active) and reference conductors by short trapezoidal pulse. Rise, top and fall times for the pulse are equal to 100 ps each, while the magnitude on matched load is equal to 500 V. Waveform of the voltage between signal and reference conductors at the end of 6 cascaded sections is shown in Fig. 7. It is seen that an original pulse is decomposed to  $2^6=64$  pulses. They must have magnitudes  $500/64=7.8$  V, but the observed magnitude is about 7.5 V, possibly, due to small mismatching. (It was assumed when modeling that losses and dispersion in lines are negligible.)

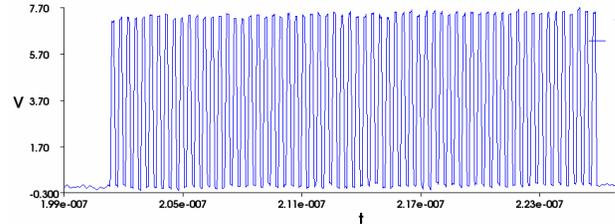


Fig. 7. Resulting waveform (V, s) of 500 V input pulse decomposition in printed modal filter into 64 pulses with small magnitudes (by factor of 64).

To obtain more proper  $R$  value, an analytic expression for even/odd mode magnitudes (normalized to  $E$  value) is used  $U_{e,o}/E = (1+pL_{e,o})/2 \cdot P_{e,o}$  (1) where  $pL_{e,o} = (R-Z_{e,o})/(R+Z_{e,o})$ ,  $P_{e,o} = 1+R/Z_{e,o}$ .

Equalization of even/odd mode magnitudes after simple manipulations gives the well known condition  $R=(Z_e Z_o)^{1/2}$ . (2)

Pulse magnitude for  $R=(Z_e Z_o)^{1/2}$  is easy to obtain analytically, substituting (2) into (1), after simple manipulations as  $U/E=(Z_e/Z_o)^{1/2}/((Z_e/Z_o)^{1/2}+1)^2$ . (3)

This simple, but general formula expresses essentially the dependence of the one-stage MF attenuation on the coupling through the relation of even and odd mode impedances. Use of analytic formula (3) permits quick and accurate estimation of MF attenuation without computation of time domain response.

On condition (2) the voltage at the input of the filter is  $U_0=E/2$ . Then, assigning the output voltage of a stage by  $U_1$ , we have one-stage attenuation

$$U_1/U_0=2(Z_e/Z_0)^{1/2}/((Z_e/Z_0)^{1/2}+1)^2, \quad (4)$$

and assigning  $k=(Z_e/Z_0)^{1/2}$ , we have simpler forms for 1 stage

$$U_1/U_0=2k/(k+1)^2 \quad (5)$$

and  $n$  stages

$$U_n/U_0=[2k/(k+1)^2]^n. \quad (6)$$

The dependency of the power dissipation on the resistors of the  $n$ -stage MF on the input pulse duration  $t_{in}=t_d+0,5t_r+0,5t_f$  (at the level of 0.5 of the magnitude) has been investigated and the following results have been obtained:

1. The power at the beginning of the passive conductor is distributed among the stages in proportion to their length.
2. The sum of average powers at the beginning of the passive conductor of all stages remains almost constant for MF implementations with any number of stages provided that the total mode delay difference is the same in all MF implementations.
3. The more stages are implemented in the MF, the bigger total power dissipates on the loads at the end of the passive conductor of all stages. This trend can be observed for all ranges under study (in the range as well as out of the range of MF effective filtration).
4. Increase in the number of the MF stages leads to the decrease of the MF effective filtration range since the input pulse energy distributes between the active and passive conductors equally.
5. If the stage length grows from the beginning to the end of the MF, the power which is dissipated on the resistors at the beginning of stages grows by factor of 2 from stage to stage; the power which is dissipated on the resistors at the end of stages decreases by factor of 2 from stage to stage.

### III. PRACTICAL RESULTS

#### A. Improved Simulation

The implementation of the improved iterative methods for solution of linear systems decreased the time needed for the EMC simulation of spaceborne equipment. The investigation of the spline approximation and the Godunov's method lead to the development of a universal instrument for approximation of different dependencies and alternative approach to the calculation of transmission line (nonregular and with non-linear loads) time-domain response. The usage of the alternative approach is useful for verification of the simulation results obtained by means of the main approach.

The application of the Godunov's method to such structures has been done for the first time. Numerical method of inverse Laplace transformation has been implemented as well. Response calculation results for structures of 1, 2, 3 line sections with linear and non-linear loads have been presented. To demonstrate the usage of the implemented methods, examples of the analysis of real electric interconnections in spaceborne equipment PCBs have been obtained.

Capacitances for different real 3D-structures of spaceborne equipment have been calculated: contact pads, footprints, crossing of two conductors, interlayer via, SNP 339 connector.

Analytic models for shielding effectiveness (SE) calculation of typical structures (metal plate and rectangular chassis with a

slot) suitable for quick estimations of real spaceborne equipment structures have been implemented. These models have been tested on the following comparative estimations: magnesium alloy and aluminum plates, UEM case of various sizes, connector case with an overlapping slot. Example of electric field SE frequency dependence calculated for a distance of 1 mm from the metal plate is shown in Fig. 8(a). One can see a presence of the pronounced SE minimum, because of which the SE is reduced at some frequency range with increasing the frequency. For copper, the minimum corresponds to the frequency of 0.1 MHz, for aluminum – 0.2 MHz, and for magnesium alloy – 1 MHz. It is the presence of this minimum that leads to the existence of the frequency range on which the SE degradation increases with increasing the frequency. For example, the SE at the frequency of 0.2 MHz is reduced by 20 dB, whereas at 1 MHz – 30 dB. It is worthy of note that this minimum dropouts with the plate thickness increasing, as shown in Fig. 8(b). Therefore, similar estimations are relevant also for a thin foil.

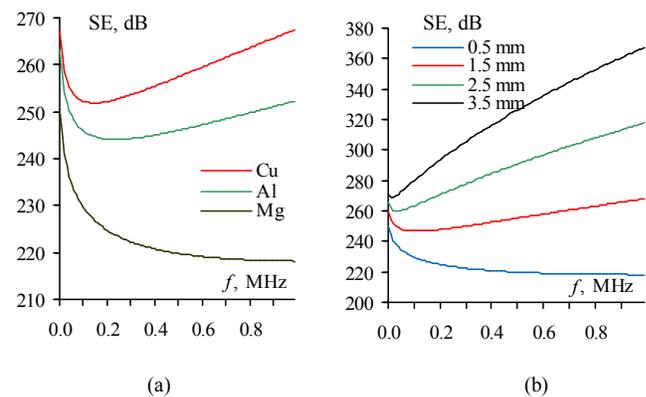


Fig. 8. Frequency dependence of electric field SE at a distance 1 mm from a plate: (a) of thickness 0.5 mm for different metals; (b) of thickness 0.5–3.5 mm for magnesium alloy

Example of SE connector case (Fig. 9) estimation, being actual for protection of sensitive circuits' junctions for a spacecraft with unpressurised chassis from external fields, is shown in Fig. 10. It shows the influence of the aperture ( $w$ ) of the front wall of the connector case ( $a \times b \times d = 29.5 \times 8 \times 21.5$  mm) in the frequency range up to 20 GHz. From Fig. 10 one can see that at frequencies up to 1 GHz for the gap aperture of 2 mm SE value increases about 20 dB in comparison with a fully open aperture, while in the intervals between the resonant frequencies it increases about 10 dB, whereas at the resonance frequencies the SE may degrade.

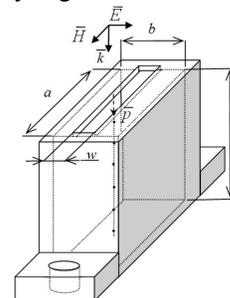


Fig. 9. A model of connector case

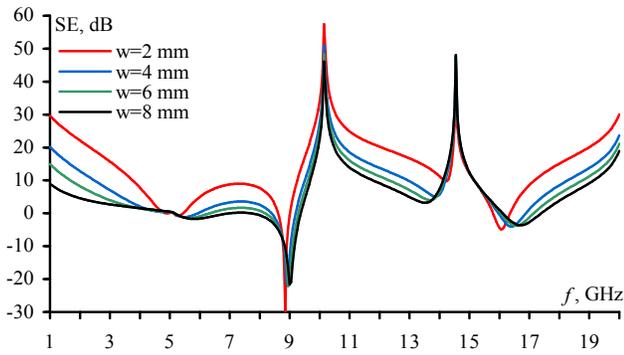


Fig. 10. Frequency dependence of SE inside of the connector case at distance  $p=10$  mm

**B. Modal Technologies**

A technology for reduction of signal distortion in the printed conductors using optimal choice of conductor and dielectric parameters has been proposed. To decrease modal distortion in active conductor and far-end crosstalk in structures with nonhomogenous (in cross-section) dielectric filling, an additional dielectric layer (such as moisture-protective coating layer) covering the PCB surface has been proposed. The applicability of this technology has been verified by simulation of a real PCB fragment. The value of this technology consists in its implementation without either change of routing or introduction of additional components. Only the optimal varnish thickness must be selected at the latest stage of the PCB manufacture. This can be easily done, for example, by means of the polyparaxylylene covering with precisely controlled thickness. The described technique is demonstrated by the simulation of the eight-wire bus of the spacecraft PCB (Fig. 11).

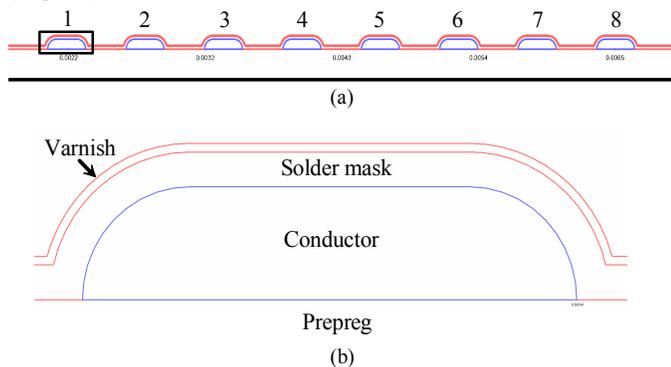
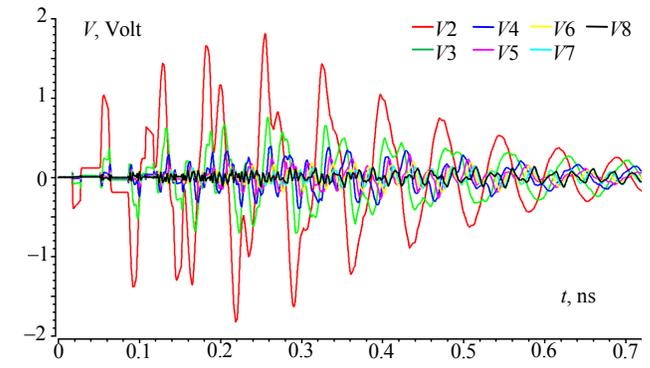
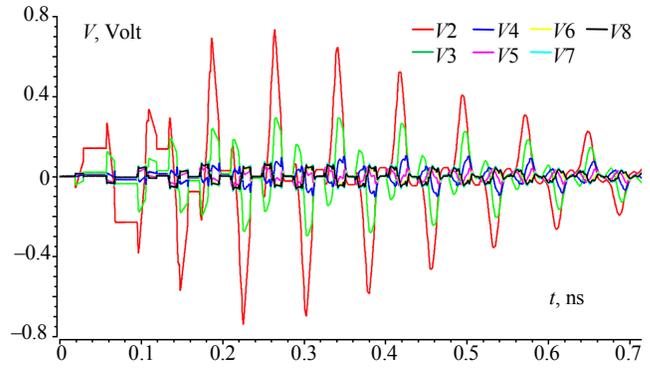


Fig. 11 Eight-wire bus: (a) cross section and (b) increased fragment of it

Thickness of a covering varnish  $hV$  is changed with fixed other parameters. The near end of the outer conductor was excited by the trapezoidal pulse (EMF of 6 V, front/decay duration of 1 ns, flat top duration of 8 ns). The crosstalk waveforms at the ends of passive conductors of the bus are shown in Fig. 12 (for case of  $5 \Omega$  at the near ends and  $1 \text{ M}\Omega$  at the far ends of passive conductors) for different  $hV$  values. One can observe the original crosstalk level of 2 V for  $hV=5 \mu\text{m}$  and the crosstalk reduced down to 0.8 V for  $hV=100 \mu\text{m}$ .



(a)



(b)

Fig. 12. Far-end crosstalk waveforms of eight-wire bus for varnish thickness of (a) 5 and (b) 100  $\mu\text{m}$

A modal filtration technology has been developed and patented [17] which is based on a new principle of protection against short pulses with dangerously high voltage in transmission lines. It uses the phenomenon of pulse signal decomposition in cascaded transmission line sections. Application of this technology allows improving the conducted susceptibility of power and signal circuits of PCBs with “system-on-a-chip” components. Analysis and implementation of modal filtration in various elements of spaceborne equipment yielded the following results: the manufacture of PCBs with integrated modal protection has been proposed; calculation of MF characteristics for nominal values of PCB fiber glass and foil thicknesses; MF calculation sequence has been presented; possible options for structural implementation of MF have been proposed; an experimental investigations of ultra-short pulse propagation in 6 prototypes of single-stage MF with the broad side coupling (as the most promising), frequency characteristics of 2 prototypes as well as ultra-short pulse propagation in MF based on the flat cable have been performed. A general technique for separate and joint (with traditional protection methods) usage of modal filters has been presented [18]. The application of the steps of the technique for creation of MF based on the flat printed cable or of the printed MF has been shown. Design documentation has been developed for printed MF production.

To illustrate this results a symmetrical structure of single-stage MF with the broad side coupling is shown in Fig. 13, where A, P, R denote active, passive, and reference conductors, accordingly, while  $R$  is resistance value.

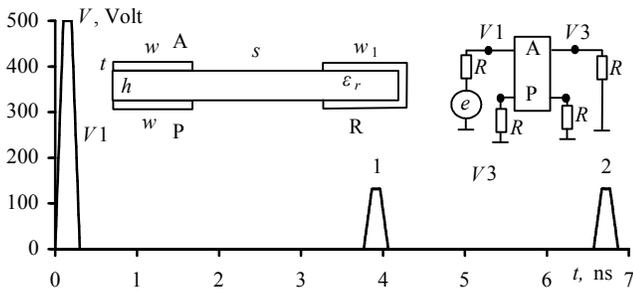


Fig. 13. Cross section of symmetrical modal filter structure, schematic of the connections and waveforms of the voltage at the active conductor

The following features of this structure are observed. The active and passive conductors and reference conductor as well are axially symmetric. Width of conductors ( $w, w_1$ ) may be increased for high current, while the separation of conductors ( $s$ ) may be used in order to keep the defined value of characteristic impedance. However, the main advantage of this structure concludes in the fact that the odd mode propagates mainly in dielectric substrate, while the even mode propagates only partly in dielectric substrate, but considerably – in air.

Cross section of other (asymmetric) structure of MF is shown on Fig. 14. Peculiarity of this structure is the absence of the U-shaped strip line that simplifies MF placement on the PCB. As the result the excitation pulse magnitude is decreased by factor 5 for the difference of modal delays of 3 ns/m with MF length of 0.2 m.

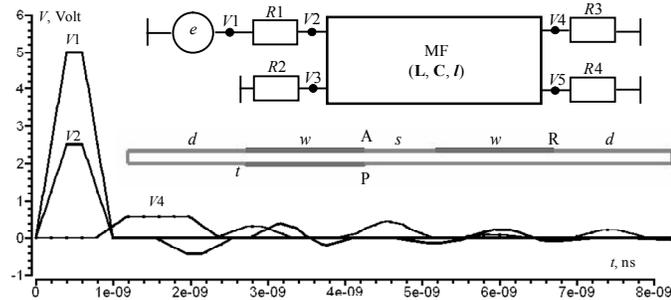


Fig. 14. Cross section of asymmetrical modal filter structure, schematic of the connections and waveforms of the voltage at the active conductor

A technology of modal decomposition and restoration, or “modal excitation” in short, has been proposed and patented [19] in order to reveal hidden possibilities of modal decomposition and restoration of pulse and harmonic excitations with dangerous magnitude in spaceborne equipment structures.

An essence of the modal decomposition and restoration phenomenon consists in the following. If the protective equipment (PE) is included between a signal and reference conductors before the protected circuit, the dangerous pulse may decompose into pulses with smaller magnitude at the end of section 1 (Fig. 15). As the result the PE will not protect if the amplitude of the decomposed pulses is below the threshold of the PE sensitivity. Moreover at the end of section 2 restoration of dangerous pulse can happen because the modes simultaneously come at the end of section 2. As the result it may destroy the functioning the subsequent circuits.

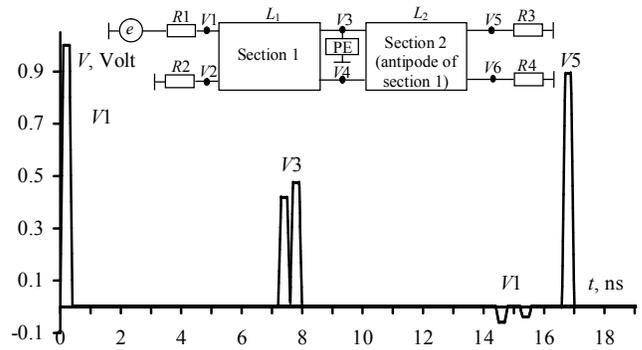


Fig.15. Schematic of the connections and waveforms of the voltage

The modal antipodes are the segments of coupled lines wherein the per unit of length propagation delays of odd and even modes are equal in magnitude but opposite in sign. Since the difference of the modal delays of coupled lines is determined by its length multiplied by difference of the per unit of length propagation modal delays it is last is determining the sign of the difference of the modal delays. Therefore, this is important to investigate in future the modal antipodes for various real structures.

The modal probing technology has been proposed for contactless detection, identification and diagnostics of multiconductor interconnects. Two implementation options (active [20] and passive [21]) have been patented. Under detection the ability to detect passive (probed) conductors is meant. Under the identification the ability to determine the amount of probed conductors is meant. Under diagnostics the ability to determine passive (probed) conductors breaking is meant. It is known that the pulse signal in the  $N$ -conductor (not including the reference) line with an inhomogeneous dielectric filling may be subjected to modal distortion up to modal decomposition on  $N$  pulses of smaller amplitude due to differences of modal delays. Complete decomposition of pulse signal in the line with length  $l$  will occur if the total pulse duration  $t_\Sigma$  less than the minimum modulus of the difference mode delay propagation in structure, i.e. under the condition  $t_\Sigma < l \cdot \min|\tau_i - \tau_k|$ ,  $i, k=1, \dots, N$ ,  $i \neq k$ ,

where  $\tau_{i(k)}$  – delay per unit of length for  $i(k)$ -th mode of structure. This phenomenon can be applied for detection, identification and diagnostics of multiconductor interconnects. Generalization of these opportunities herein is called the modal probing. If probed conductors have different electrical and magnetic couplings with the probing line, the information about probed conductors can be obtained from the form of modal signal distortion in the probing line.

Block diagram of the device that implements the principles of modal probing is shown in Fig. 16. The device operates as follows: the probing pulse from the generator output goes to a probing line. The probing pulse passing along the probing line is subjected to modal distortion caused by the presence of probed wire structure. The receiver gets the signals from the input and output of the probing line and sends information to the processing unit. All units of the device function according to signals of control unit. From the shape of the signals at the

near and far ends of the probing line information about the probed structure is obtained.

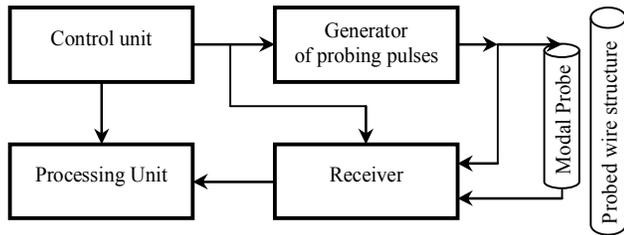


Fig. 16. Block diagram of the active modal probing device

The possibility of detection and identification of electrical interconnections by modal probing is illustrated by quasi-static modeling of the trapezoidal pulse (keystone-shaped) signal distortion in the microstrip structures 1.5 m in length (Fig. 17). For  $N=2$  (Fig. 17a) at the far end of the probing line ( $V3$ ) there are two pulses rather than one. The second pulse was caused by the presence of the probed passive wire (and, as a consequence, by the excitation of even and odd modes), electric and magnetic couplings with the probing line, and by the fact that the total duration of the input pulse was less than the total difference between modes delays. The difference of modes delays is due to the inhomogeneous dielectric filling of structure. For  $N=3$  (Fig. 17(b)) at the far end of the probing line ( $V3$ ) there are three pulses rather than two. The appearance of three pulses is caused by the presence of two passive conductors, so three modes are excited in the structure and the delay difference between them is more than the pulse duration.

Thus, these results show that using the number of pulses at the far end of the active conductor one can determine the presence and amount of passive conductors i.e. solve the problem of detection and identification of electrical interconnections.

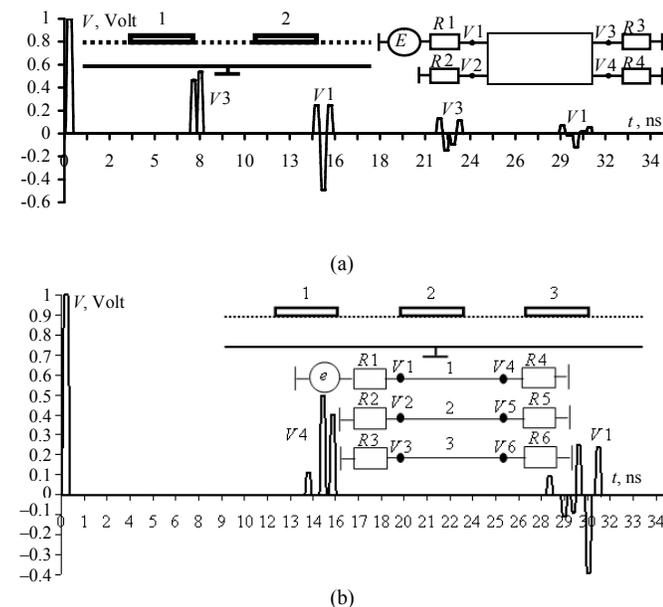


Fig. 17. The waveforms at the input and output of active microstrip line when (a)  $N=2$  and (b)  $N=3$

The possibility of diagnostics is illustrated by results of simulation and experiment of the pulse propagation along the flat cable of the PUGNP  $3 \times 1.5$  type (Fig. 18a).

For the diagnostics of a passive wire by the modal probing the form of a modal distortion of the pulse signal should vary depending on the condition of passive wire. One of the important problems of electrical connections diagnostics is to determine the wire breaks. Let's consider in Fig. 18(b) an effect of passive wire break to a form of modal distortion of the pulse signal in the probing line. The waveform at the far end of the probing line is shown in Fig 19. As can be seen, when passive wire is break, four pulses come to the far end of the probing line instead of two as in the case without break. (Partial overlap of the pulses is due to dispersion.) Waveforms at the far end of the probing line under various boundary conditions at the ends of the passive wire are considered in more detail in [22]. Modal distortion in frequency domain is considered in [23].

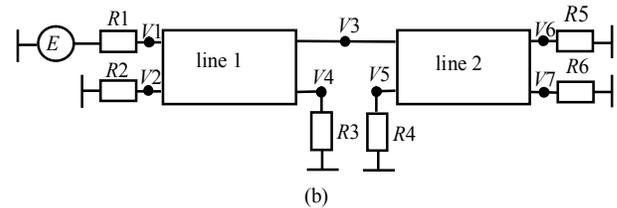
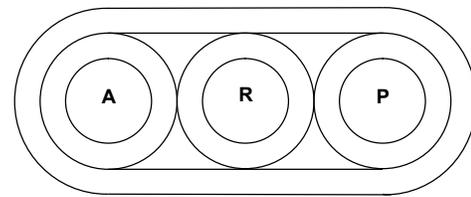


Fig. 18. Cross section (a) of the  $3 \times 1.5$  flat cable of the PUGNP type and (b) schematic diagram of the examined structure with break in passive wire

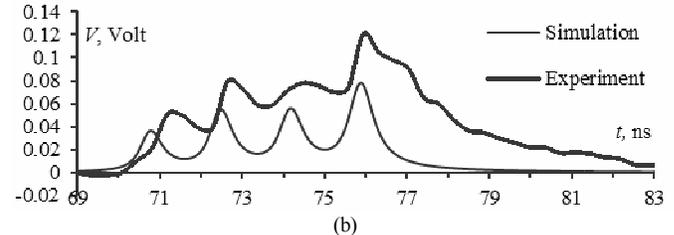
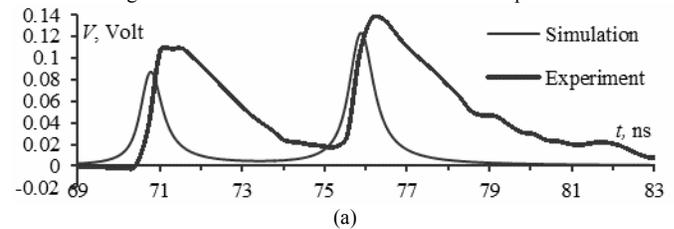


Fig. 19. The waveforms at the far end of flat cable (a) without and (b) with passive wire break

### C. Environmental shielded TEM chamber

At the present time the measurement systems for EMC testing, based on various methods and equipment are actively improving. One way to improve the characteristics and to enhance the tests conducted using various equipment is their hybridization. For this propose the idea of integration of TEM-

cell into the environmental shielded chamber is proposed and actively developed. It will permit to gain new knowledge about the interaction between internal and external electromagnetic and climatic effects on a TEM-cell and a device under test (DUT), located in the internal volume of the TEM-cell. Besides, it will help to bring the EMC tests (measurements of emission and immunity) to the real operating conditions of the DUT. For practical purposes, it will be possible, for example, to identify failure mechanisms in semiconductor components subjected to the thermal and electromagnetic fields. This is important when the component is operating at limits of the temperature range where the risk of failure of its p-n junction due to the cumulative effects of various factors (including an ultra wide band pulse) raises. The above-described concept is implemented in a new type of chamber (Fig. 20).

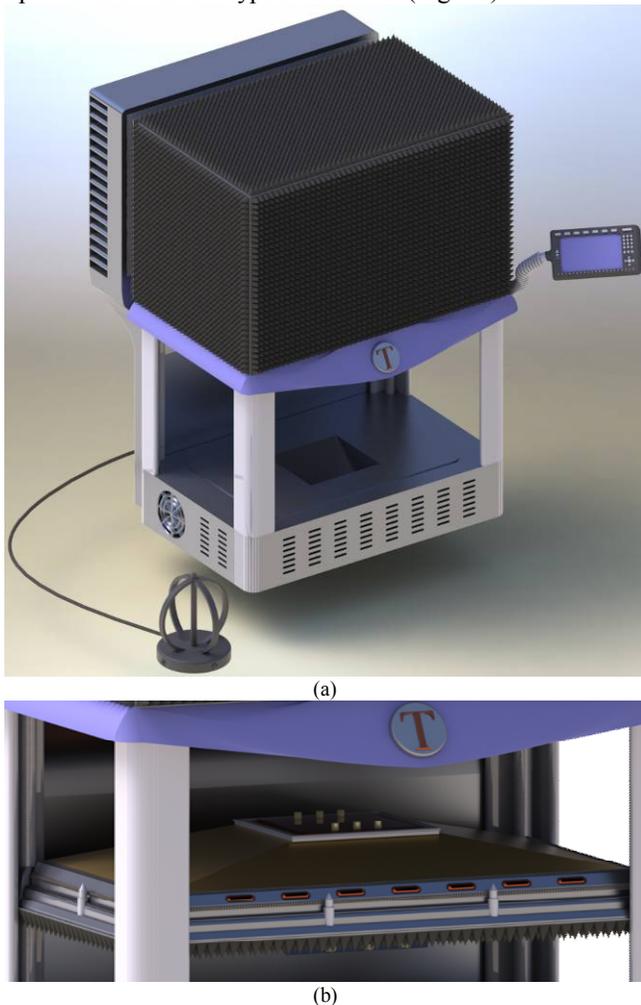


Fig. 20. Environmental shielded TEM chamber: (a) exterior, (b) view with the open door

The chamber design shown in Fig. 20(a) includes an outside case with multilayer shield and RF absorber; external sensor of temperature and electromagnetic field; MPU control system; LCD display for digital and graphic visualization of the temperature and humidity inside the TEM-cell, automated removable door at the bottom shown in Fig. 20(b), on the inner side of which a test table is placed for the DUT, occurring inside the TEM-cell after closing (lifting up) the door. The

chamber will allow for research and EMC testing of a component or a small device on single and joint effects of temperature and humidity, as well as measurements of emissions from it.

Also, with this chamber one can get more advanced Spice, IBIS and ICEM models of components. Particularly, research and testing of new components (modal filters) family is planned in this chamber.

The methods of component tests are based on EMC standards, in particular on emission and immunity tests of IC, and also include common standards for EMC testing of radioelectronic equipment in TEM-cells. During these tests the DUT characteristics are measured in wide frequency range (up to 3 GHz) when exposing to electric field (up to 3 kV/m) with temperature ( $-50$  to  $+150^{\circ}\text{C}$ ) and humidity (up to 90%). SE of the chamber case is more than 40 dB at frequencies up to 40 GHz.

#### D. Other Results

The optimization of spaceborne equipment by EMC criteria (EMC optimization in short) is characterized by a large number of parameters and local optima of the objective function in conjunction with resource-intensiveness and diversity of the underlying analysis problems. Even the application of evolutionary algorithms which is the most appropriate option for such optimization problems becomes increasingly complicated as the complexity of the prospective space vehicle EMC simulation grows. Therefore, it is necessary to enhance the existing optimization algorithms and propose techniques for their effective usage in practice. To this end, a technique for EMC optimization of spaceborne equipment has been developed and its main steps have been described. The technique has been tested on several examples of parameter optimization: a microstrip line, a wide-band mathematical model of resistor including parasitics, a multiconductor transmission line as well as structural optimization of multi-stage MF.

Preliminary simulation of EMC testing in accordance with MIL-STD-461F standard has been implemented in the part of conducted emissions from power circuits of the “system-on-a-chip” components.

Preliminary measurement results of the reflection ratio  $S_{11}$  in the frequency range from 10 MHz to 20 GHz have been presented for two typical components (capacitor and resistor). A significant difference between characteristics of idealized elements, classical model and real components has been demonstrated.

Below the creating the models of the SMD resistor and the wired capacitor is briefly described. For this aim the following technique was used: 1) measurement of reflection coefficient  $S_{11}$  frequency dependence for a component; 2) calculation of impedance  $Z$  from  $S_{11}$ ; 3) approximation of the  $Z$  frequency dependence by a rational function; 4) expansion of the rational function on partial fractions; 5) realization of partial fractions by the equivalent circuits, using the methods of circuit synthesis; 6) generation of total SPICE-model; 7) model verification.

Measurement of  $S_{11}$  is performed on the vector network analyzer. For performing the measurements the components are soldered on SMA connector of coaxial-microstrip transition type [24]. To reduce the transition influence the connector pin was previously shortened.

After calculation of  $Z$  from  $S_{11}$  the approximation of  $Z$  by a rational function using the vector fitting method [25] is performed. As a result the following rational function is obtained

$$f(s) = \sum_{m=1}^N \frac{c_m}{s - a_m} + d + se \cdot \quad (7)$$

The first member of the function is decomposed into several sums of fractions with complex conjugate poles of the form

$$\frac{a + jb}{s - (\sigma + j\omega)} + \frac{a - jb}{s - (\sigma - j\omega)} \quad (8)$$

Each such sum (8) is realized as a parallel  $\pi$  circuit (Fig. 21), but the remaining members – as cascaded resistance and inductance. To calculate the parameters of the circuit for the resistance and inductance the formula from [26] are used. Total equivalent circuit for the resistor is consisting of 4 cascaded parallel circuits, resistance and inductance and for the capacitor – of 6 circuits.

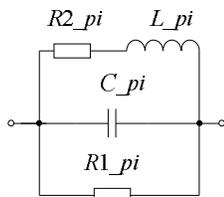


Fig. 21. Schematic of the parallel circuit

For verification of the models their circuit analysis in TALGAT system was made. The frequency dependences of  $Z$ , calculated using the obtained models of the resistor and capacitor are shown in Fig. 22.

Root mean square deviations (RMS) between results of  $|Z|$  measurement and modeling are equal to  $0.701 \Omega$  for the resistor, and  $11.05 \Omega$  for the capacitor. Thus, it is possible to conclude that the obtained models are exact and suitable for PCB EMC analysis.

To import real PCB configurations and prepare them for EMC simulation, a TLPCB conversion module has been created. Commands implemented in the TLPCB module allow reading the PCB designed in Altium Designer into the memory and convert it into a format suitable for further processing using the TALGAT software. Example of these automatic transformations is demonstrated in Fig. 23, wherein firstly a net  $A-B$  is chosen with defined distance ( $s$ ) to account for presence of neighboring conductors and cross sections of the resulting structure are obtained consequently at distance of  $d$ , then 6 intervals ( $len1-len6$ ) having the unique sections are revealed using a graph theory, at last, circuit diagram consisting of cascaded multiconductor transmission lines described by according  $L, C, R, G$  matrixes is obtained. One can observe that mutual couplings of neighboring conductors can be properly and automatically considered.

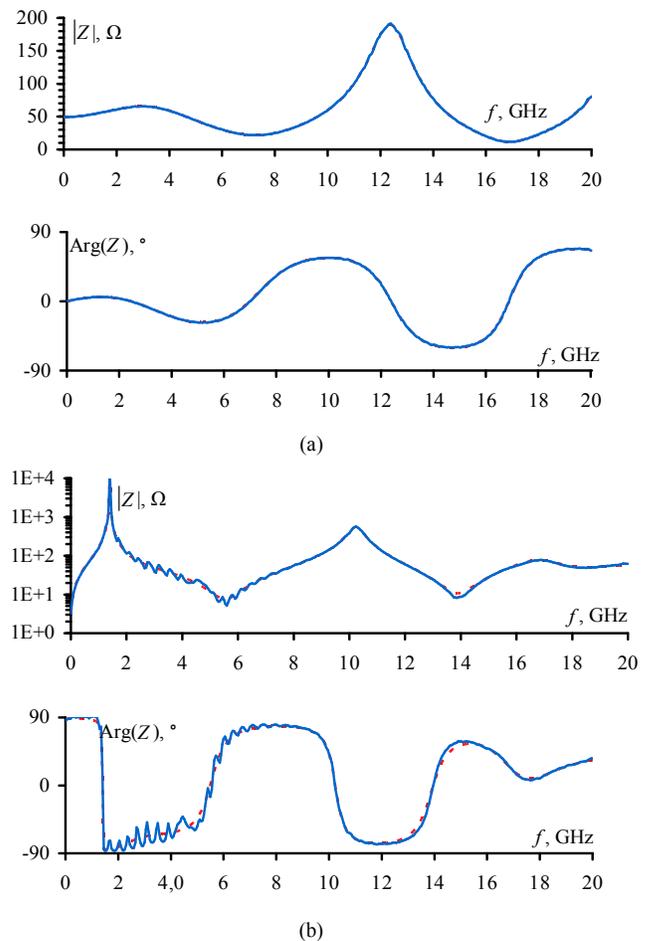


Fig. 22. Frequency dependence of impedance  $Z$  for (a) resistor and (b) capacitor: measurement (—), modeling (---)

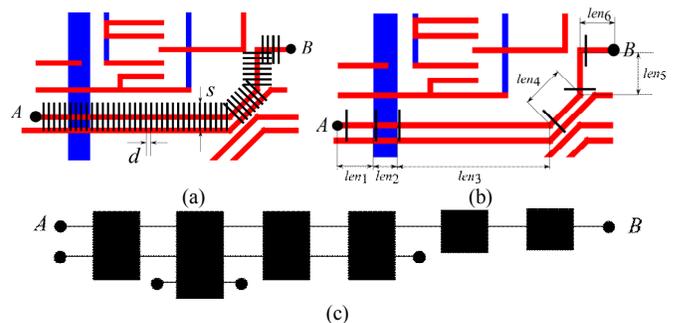


Fig. 23. Automatic transformations of (a) printed nets to (b) line intervals and (c) to a circuit diagram

As a further extension of this PCB import the capability of more detailed representation of the wide bus bending is implemented for more precise simulation of high speed signals propagation. It is schematically shown in fig. 24.

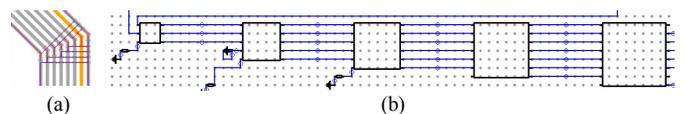


Fig. 24. Sketch of (a) bus bending and of (b) a half of its precise circuit presentation

A technique for preliminary EMC analysis of PCBs based on qualitative analysis (without simulation) has been developed

which allows for obtaining of specific recommendations for improvement of the EMC of PCBs without knowing details of PCB circuitry engineering and just by changing the PCB routing. The technique has been tested by preparation of preliminary recommendations for 5 real PCB. All recommendations have been taken into account by designers, and according to a number of recommendations (about 100), changes in the PCB routing have been introduced. Similarly, recommendations for a more complex PCB (radio navigation electronic equipment unit of spacecraft) have been developed.

Techniques for obtaining of Spice-models have been developed for the following analog components: capacitor, resistor, diode, bipolar transistor. Technique for creation of models has been developed which takes into account parasitic parameters. For some integrated circuits (IC) and analog components the preliminary models have been created. Techniques for creation of IBIS-models based either on physical measurements or Spice-model conversion have been developed. An example of database for ICs used for spaceborne equipment has been created.

The results of the improvement of the technique for preliminary EMC analysis of PCBs have been presented for routing of back-up circuits. The proposed approach for reduction of radiated emissions has been shown to be effective for the open structure as well as in the presence of conductive plate and shielding box.

The results of the usage of simulation for preparation of recommendation to ensure the EMC have been presented for PCB power-ground circuits. A possibility to decrease the inductance of power-ground circuits by factor of 9 by means of conductor routing has been shown.

The following programs and techniques have been developed for real PCBs: programs and technique for simulation of electric circuit which take into account parasitic parameters of components and PCB; a technique for search for circuit parts with problems; a technique for signal integrity analysis of real PCB in Altium Designer; a technique for post-topological analysis of crosstalk in Altium Designer; a technique for signal integrity analysis using TALGAT software; program and techniques for electromagnetic simulations.

#### IV. CONCLUSIONS AND OUTLOOK

In the paper, the results of the recent EMC projects have been presented and summarized. First, the outcome in the field of quasi-static and electromagnetic analysis has been described, in particular, models for calculation of time-domain response and capacitive matrix. After that, improvements for solution of linear algebraic systems and corresponding speed-ups (up to 12 times) have been discussed. Then modal filtration in PCBs has been considered.

The practical significance of the obtained results has been highlighted in the second section where details about developed methods for improved EMC simulation have been provided, including the spline approximation, the Godunov's method, the analytic models for SE calculation of typical structures. Technologies for reduction of signal distortion,

modal filtration, modal decomposition and restoration and modal sensing have been presented with corresponding references. Finally, proposed techniques for EMC improvement of spaceborne equipment have been briefly presented. A number of theoretical and practical results obtained are used and will be implemented in the new project on the development of prospective equipment for global navigation satellite system. The simulation of this equipment will be performed using the TALGAT software which includes models and features developed during the past project and modified during the new project. Application of the techniques presented in the paper will guarantee high functional and EMC characteristics of the prospective equipment. At last the obtained results have formed a base for their extensions in the current and future EMC projects of TUSUR University.

#### REFERENCES

- [1] T. Gazizov, A. Melkozerov, A. Zabolotsky, P. Orlov, V. Salov, R. Ashirbakiev, R. Akhunov, S. Kuksenko, I. Kalimulin, "New results on EMC simulation for space projects of TUSUR University," Proc. of IEEE Int. Conf. on Numerical Electromagnetic Modeling and Optimization for RF, Microwave, and Terahertz Applications. May 14–16, 2014, Pavia, Italy.
- [2] T.R. Gazizov and N.A. Leontiev, "Transient response of a periodic transmission line structure with capacitively loaded junctions," Proc. of the 1997 Sino-Japanese Joint Meeting on Optical Fiber Science and Electromagnetic Theory. October 14–16, 1997, Wuhan, China, pp. 322–327.
- [3] T.R. Gazizov, "Analytic expressions for Mom calculation of capacitance matrix of two dimensional system of conductors and dielectrics having arbitrary oriented boundaries," Proc. of the 2001 IEEE EMC Symposium, Montreal, Canada, August 13–17, 2001, vol. 1, pp. 151–155.
- [4] T.R. Gazizov, "Calculation of capacitance matrix of three dimensional multiconductor system in multiple dielectric media," Record of International Symposium on Electromagnetic Compatibility. Magdeburg, Germany, October 5–7, 1999, pp. 31–36.
- [5] S.M. Rao, D.R. Wilton, A.W. Glisson, "Electromagnetic scattering by surfaces of arbitrary shape," IEEE Transactions on antennas and propagation, May 1982, vol. 30, no. 3, pp. 409–418.
- [6] I.S. Kostarev, T.R. Gazizov, Yu.M. Kazantsev, "Analytic evaluation of the matrix entries for linear algebraic systems in the problem of electromagnetic scattering by surfaces of arbitrary shape," Journal of Mathematical Sciences, Vol. 199, no. 4, June, 2014, pp. 456–462.
- [7] R.R. Akhunov, S.P. Kuksenko, V.K. Salov, T.R. Gazizov, "Multiple iterative solution of linear algebraic systems with a partially varying matrix," Journal of Mathematical Sciences, Vol. 199, no. 4, June, 2014, pp. 381–385.
- [8] V.K. Salov, T.R. Gazizov, O.A. Nikitina, "Convergence of multiple iterative solution of linear algebraic systems with a fully varying matrix using a single calculated initial preconditioner," Innovative Information Technologies: Materials of the International scientific-practical conference. Part 2. / Ed. Uvaysov S.U.–M.: HSE, 2014. April 21–25, 2014, Prague, Czech – P. 452–457.
- [9] R.R. Akhunov, S.P. Kuksenko, V.K. Salov, T.R. Gazizov, "Sparse matrix storage formats and acceleration of iterative solution of linear algebraic systems with dense matrices," Journal of Mathematical Sciences, vol. 191, May, 2013, pp. 19–27.
- [10] R.R. Akhunov, S.P. Kuksenko, V.K. Salov, T.R. Gazizov, "Optimization of the ILU(0) factorization algorithm with the use of compressed sparse row format," Journal of Mathematical Sciences, vol. 191, May, 2013, pp. 19–27.
- [11] T.R. Gazizov, A.M. Zabolotsky, I.E. Samotin, "Modal decomposition of UWB pulse in power cable structures: simple experiment showing useful possible applications," Book of abstracts EUROEM 2008, 21–25 July 2008, Lausanne, Switzerland, p. 62.
- [12] T.R. Gazizov, I.E. Samotin, A.M. Zabolotsky, A.O. Melkozerov, "Design of printed modal filters for computer network protection," Proc.

- of 30-th Int. conf. on lightning protection, Sept. 13–17, 2010, Cagliari, Italy, pp. 1246-1–1246-3.
- [13] T.R. Gazizov, A.M. Zabolotsky, A.O. Melkozerov, E.S. Dolganov, P.E. Orlov, “Improved design of modal filter for electronics protection,” Proc. of 31-st Int. conf. on lightning protection, Sept. 2–7, 2012, Vienna, Austria, pp. 1–4.
- [14] T.R. Gazizov, A.M. Zabolotsky, A.O. Melkozerov, E.S. Dolganov, P.E. Orlov, “Analysis of power dissipation in resistive terminations of single- and multistage modal filters,” Proc. of 31-st Int. conf. on lightning protection, Sept. 2–7, 2012, Vienna, Austria, pp. 1–4.
- [15] T.R. Gazizov, A.M. Zabolotsky, “Experimental results on UWB pulse propagation in low-voltage power cables with different cross sections,” IEEE Transactions on electromagnetic compatibility, vol. 54, no. 1, February 2012, pp. 229–231.
- [16] T.R. Gazizov, A.M. Zabolotsky, I.E. Samotin, A.O. Melkozerov, “Simple and free mitigation of short pulse lightning effects by flat power cables,” Proc. of 30-th Int. conf. on lightning protection, Sept. 13–17, 2010, Cagliari, Italy, pp. 993-1–993-3.
- [17] Patent of Russian Federation №2431897.
- [18] T.R. Gazizov, A.M. Zabolotsky, A.O. Melkozerov, P.E. Orlov, I.G. Bevenko, E.S. Dolganov, “Evaluations of protection methods using TVS-array and modal filter,” Book of abstracts EUROEM 2012, 2–6 July 2012, Toulouse, France, p. 106.
- [19] Patent of Russian Federation №2431912.
- [20] Patent of Russian Federation №2386964.
- [21] Patent of Russian Federation №2456588.
- [22] P.E. Orlov, T.R. Gazizov, A.M. Zabolotsky. Experimental confirmation of the possibility for contactless diagnostics of multiconductor structures using modal probing // Russian Physics Journal. November 2013, Volume 56, Issue 6, pp 652–656.
- [23] P.E. Orlov, T.R. Gazizov, A.M. Zabolotsky. Frequency Analysis of Modal Distortions and its Application to Diagnostics of Electric Connections // Russian Physics Journal. January 2014, Volume 56, Issue 9, pp 1099–1101.
- [24] I.F. Kalimulin, T.R. Gazizov, A.M. Zabolotsky “Impedance of low-frequency passive components of spaceborne equipment at frequencies ranging to 20 GHz,” Instruments and Experimental Techniques, vol. 55, no. 2, pp. 231-237, 2012.
- [25] B. Gustavsen, “Improving the pole relocating properties of vector fitting,” IEEE Trans. Power Deliv., vol. 21, no. 3, pp. 1587–1592, 2006.
- [26] N. Balabanian, Network Synthesis. Englewood Cliffs: Prentice-Hall, 1958, p. 440.

# The first principles study on the TbP compound

Y.O. Ciftci<sup>A\*</sup>, Y. Mogulkoc<sup>B</sup> and M. Evecen<sup>C</sup>

**Abstract**— The structural, elastic, electronic, thermodynamic and vibrational properties of TbP which crystallize in NaCl (B1), CsCl (B2), ZB (B3), tetragonal (L10), WC (Bh), NiAs (B8), PbO (B10) and wurtzite (B4) structures were analyzed by performing ab-initio calculations based on density functional theory using the Vienna Ab initio Simulation Package (VASP). The exchange correlation potential within the generalized-gradient approximation (GGA) of projected augmented plane-wave (PAW) was used. The calculated structural parameters, such as the lattice constant, bulk modulus and its pressure derivative and formation energy and second-order elastic constants were presented for all calculated phases. This compound exhibits crystallographic phase transition from B1 to B2 phase at pressure 55 GPa. We have performed the thermodynamics properties for TbP by using quasi-harmonic Debye model. We have also predicted the temperature and pressure variation of the volume, bulk modulus, thermal expansion coefficient, heat capacities and Debye temperatures in a wide pressure (0-50 GPa) and temperature ranges (0-2000 K) for NiAs structure. The electronic band calculations, total density of states (DOS) and partial DOS were also presented. The computed phonon dispersion curves based on the linear response method are predicted. The obtained results are compared with the available experimental studies.

**Keywords**—elastic properties, electronic properties, thermodynamic properties, TbP, structural properties

## I. INTRODUCTION

The rare-earth monpnictides have attracted the interests of many researchers due to their numerous physical properties like magnetic, elastic, thermodynamics and phonon properties [1-21]. Buschbeck et al. [1] reported on the first magnetization measurements on TbP and TbSb in magnetic fields as high as 140 kOe covering the range from

high to low temperatures. Petit et al. [2] predict an electronic phase diagram of the entire range of rare earth monpnictides and monochalcogenides, composed of metallic, semiconducting and heavy fermion-like regions and exhibiting valency transitions brought about by a complex interplay

between ligand chemistry and lanthanide contraction. Pagare et al. [3] report the high pressure behavior, electronic and elastic properties of two lutetium compounds, namely, LuAs and LuSb which crystallize in NaCl structure, by using density functional theory. J. Schoenes et al. [12] have studied optical properties of dysprosium monpnictides and presented their experimental and theoretical results. It is known that the source of these anomalous arise from the presence of 4f level close to the Fermi level [13]. Rare-earth elements are chemically very similar owing to an almost identical outer electron arrangement. Duan et al. [14] reviewed the electronic structures and magnetic properties of many rare-earth monpnictides. Because of the fully localized nature of the 4f electrons, the direct f-f interactions between neighbouring rare-earth atoms are typically considered to be closely negligible [14-16]. The earliest ab-initio electronic structure calculation of the rare-earth monpnictides was carried out by Hasegawa and Yanase in 1977 [17]. There are only some papers about TbAs and TbSb monpnictides. Nakanishi et al. [18] have investigated the Fermi surface (FS) and magnetic properties of rare-earth monpnictide TbSb by means of de Haas-van Alphen (dHvA) and high-field magnetization measurements. Nakanishi et al. [19] have investigated the magnetic and elastic properties of rare-earth monpnictide TbSb by means of specific heat, high-field magnetization, and ultrasonic measurements. Gordienko [20] has studied enthalpies of atomization and formation for some monpnictides. We have predicted structural, electronic, elastic, thermodynamic and vibration characteristics of TbN, using density functional theory within generalized-gradient (GGA) approximation [21]. The TbP compound has not been studied very intensively and deeply using the first principle methods. Terbium phosphide (TbP) is an intermetallic compound of simple rocksalt structure. After the discovery of the transition to type-II antiferromagnetism the magnetic and elastic properties of TbP have attracted considerable experimental and theoretical interest [22]. The aim of the present paper is to reveal bulk, structural properties in B1, B2, B3, Bh, L1<sub>0</sub>, B8, B10 and B4 structures and thermodynamical, electronic and elastic properties for TbP using first principles method with plane-wave pseudopotential. Method of calculation is given with some formulas in section 2. The obtained results are given with tables and figures in section 3. In last section, results and discussion are presented.

F. A. Author is with Gazi University, Department of Physics, Teknikokullar, 06500, Ankara, TURKEY (corresponding author to provide phone: +903122021266; e-mail: yasemin@gazi.edu.tr).

S. B. Author, Jr., was with <sup>2</sup>Ankara University, Department of Physics Engineering, 06100, Ankara, TURKEY (e-mail: yesim.mogulkoc @ eng.ankara.edu.tr).

T. C. Author is with <sup>3</sup>Amasya University, Department of Physics, Faculty of Arts and Sciences, 05000, Amasya, TURKEY (e-mail: meryem.evecen@amasya.edu.tr).

## II. METHOD OF CALCULATION

In the present work, all the calculations have been carried out using the VASP [23-25] based on the density functional theory (DFT). The electron-ion interaction was considered in the form of the projector-augmented-wave (PAW) method [25, 26] with plane wave up to an energy of 500 eV for B1, B2, B3 structures, 600 eV for L1<sub>0</sub>, B10 structures and 550 eV for Bh, B8 and B4 structures. This cut-off was found to be adequate for the structural, elastic properties as well as for the electronic structure. Any significant changes are not found in the key parameters when the energy cut-off is increased. For the exchange and correlation terms in the electron-electron interaction, Perdew and Zunger-type functional [27, 28] is used within the generalized gradient approximation (GGA) [26]. The  $k$ -point meshes for Brillouin zone sampling is constructed using the Monkhorst-Pack scheme [29]. The  $12 \times 12 \times 12$  for B1, B2 and B3 structures,  $11 \times 11 \times 13$  for L1<sub>0</sub> structure,  $13 \times 13 \times 14$  for Bh structure,  $14 \times 14 \times 11$  for B10 structure,  $15 \times 15 \times 11$  for B8 structure and  $13 \times 13 \times 8$  for B4 structure Monkhorst and Pack [29] grid of  $k$ -points have been used for integration in the irreducible Brillouin zone. Thus, this mesh ensures a convergence of total energy to less than  $10^{-5}$  eV/atom.

The thermodynamic properties of TbP are calculated by GIBBS program. The GIBBS code is used to investigate isothermal-isobaric thermodynamics of a compound from energy curves via quasi-harmonic Debye model [30] is used to obtain the thermodynamic properties of TbP in which the non-equilibrium Gibbs function  $G^*(V; P, T)$  takes the form of

$$G^*(V; P, T) = E(V) + PV + A_{\text{vib}}[\theta(V); T] \quad (1)$$

In Eq.(1),  $E(V)$  is the total energy for per unit cell of TbP,  $PV$  corresponds to the constant hydrostatic pressure condition,  $\theta(V)$  the Debye temperature and  $A_{\text{vib}}$  is the vibration term, which can be written using the Debye model of the phonon density of states as

$$A_{\text{vib}}(\theta, T) = nkT \left[ \frac{9\theta}{8T} + 3 \ln \left( 1 - e^{-\frac{\theta}{T}} \right) - D \left( \frac{\theta}{T} \right) \right] \quad (2)$$

(2)

where  $n$  is the number of atoms per formula unit,  $D \left( \frac{\theta}{T} \right)$  the

Debye integral, and for an isotropic solid,  $\theta$  is expressed as [31]

$$\theta_D = \frac{\hbar}{k} \left[ 6\pi V^{1/2} n \right]^{1/3} f(\sigma) \sqrt{\frac{B_s}{M}} \quad (3)$$

where  $M$  is the molecular mass per unit cell and  $B_s$  the adiabatic bulk modulus, which is approximated given by the static compressibility [32]:

$$B_s \approx B(V) = V \frac{d^2 E(V)}{dV^2} \quad (4)$$

$f(\sigma)$  is given by Refs. [31-33] and the Poisson ratio is used as 0.2070 and  $n=4$   $M= 189.904$  for TbP. Therefore, the non-equilibrium Gibbs function  $G^*(V; P, T)$  as a function of  $(V; P, T)$  can be minimized with respect to volume  $V$ .

$$\left[ \frac{\partial G^*(V; P, T)}{\partial V} \right]_{P, T} = 0 \quad (5)$$

By solving Eq. (5), one can obtain the thermal equation of state (EOS)  $V(P, T)$ . The heat capacity at constant volume  $C_v$ , the heat capacity at constant pressure  $C_p$ , the entropy  $S$  and the thermal expansion coefficient  $\alpha$  are given [34] as follows:

$$C_v = 3nk \left[ 4D \left( \frac{\theta}{T} \right) - \frac{3\theta/T}{e^{\theta/T} - 1} \right] \quad (6)$$

$$S = nk \left[ 4D \left( \frac{\theta}{T} \right) - 3 \ln(1 - e^{-\theta/T}) \right] \quad (7)$$

$$\alpha = \frac{\gamma C_v}{B_s V} \quad (8)$$

$$C_p = C_v (1 + \alpha \gamma T) \quad (9)$$

Here  $\gamma$  represent the Grüneisen parameter and it is expressed as

$$\gamma = - \frac{d \ln \theta(V)}{d \ln V} \quad (10)$$

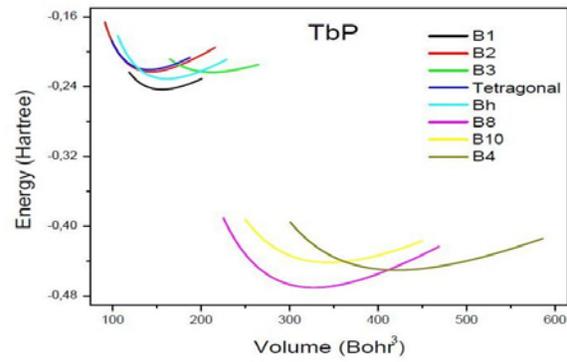
## III. RESULTS AND DISCUSSION

### Structural and electronic properties

Firstly, the equilibrium lattice parameter has been computed by minimizing the crystal total energy calculated for different values of lattice constant by means of Murnaghan's equation of state (EOS) [35] as in Figure 1. In addition, from the EOS curves as shown in Figure 1, it can be seen that in the low volume region the B8 structure is energetically favorable between investigated eight structures.

The bulk modulus and its pressure derivative have also been calculated based on the same Murnaghan's equation of state and the results are given in Table 1 along with the experimental and other theoretical values. The calculated values of lattice parameter are 5.6960 Å in NaCl (B1) structure, 3.484 Å in CsCl (B2) structure, 6.3143 Å in ZB

(B3) structure, 4.7295 Å in tetragonal (L1<sub>0</sub>) structure, 3.9145 Å in WC (Bh) structure, 4.0787 Å in NiAs (B8) structure, 5.1220 Å in PbO (B10) structure and 4.4784 Å in wurtzite (B4) structure for TbP. The NiAs (B8) structure of TbP is determined as stable phase in this study. The present values for lattice constants are also listed in Table 1 and the obtained results are quite accord with the other experimental values [36-40]. The present lattice constant in B1 structure for TbP is nearly 0.084% higher than the reference experimental values [36-40].

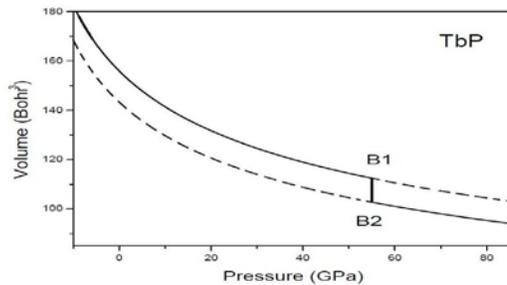


**Fig. 1.** Energy-volume curves for all calculated structures of TbP.

**Table 1.** Calculated equilibrium lattice constants ( $a_0$ ), bulk modulus ( $B$ ), pressure derivatives of bulk modulus ( $B'$ ), formation enthalpy ( $\Delta H$ ) and other theoretical works for TbP in structures

| Structure       | Reference   | a                  | c      | c/a    | B(GPa)   | B'     | $\Delta H$ (eV/atom) |
|-----------------|-------------|--------------------|--------|--------|----------|--------|----------------------|
| B1              | Present     | 5.696              |        |        | 84.2160  | 3.8253 | 2.82451              |
|                 | <i>Exp.</i> | 5.686 <sup>a</sup> |        |        |          |        |                      |
|                 | <i>Exp.</i> | 5.600 <sup>b</sup> |        |        |          |        |                      |
|                 | <i>Exp.</i> | 5.690 <sup>c</sup> |        |        |          |        |                      |
|                 | <i>Exp.</i> | 5.690 <sup>d</sup> |        |        |          |        |                      |
|                 | <i>Exp.</i> | 5.688 <sup>e</sup> |        |        |          |        |                      |
| B2              | Present     | 3.484              |        |        | 82.40567 | 3.7808 | 3.92183              |
| B3              | Present     | 6.3143             |        |        | 55.11767 | 3.6574 | 3.91157              |
| Bh              | Present     | 3.9145             | 3.5907 | 0.9173 | 76.8677  | 3.6387 | 3.48817              |
| L1 <sub>0</sub> | Present     | 4.7195             | 3.7345 | 0.7913 | 4.20908  | 3.8602 | 4.05955              |
| B8              | Present     | 4.0787             | 6.7261 | 1.6491 | 78.0407  | 3.8147 | -9.53615             |
| B10             | Present     | 5.1220             | 3.8906 | 0.7596 | 66.8960  | 3.6272 | -7.96322             |
| B4              | Present     | 4.4784             | 7.1905 | 1.6056 | 54.0484  | 3.6740 | -8.44314             |

<sup>a</sup> Ref. [36]. <sup>b</sup> Ref. [37]. <sup>c</sup> Ref. [38]. <sup>d</sup> Ref. [39]. <sup>e</sup> Ref. [40].



**Fig. 2.** Volume-pressure curves for B1 and B2 structures of TbP.

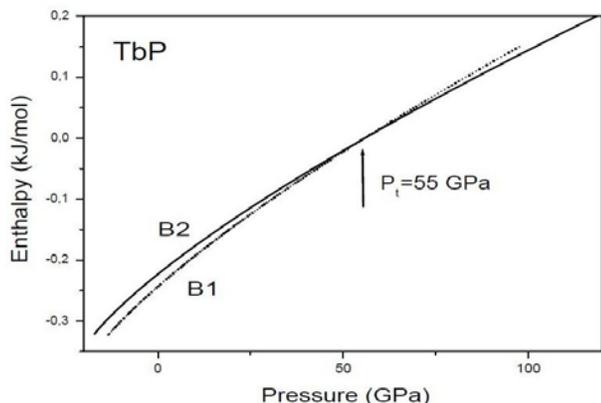
The thermodynamic stability of TbP compound in different structures can be reflected by the formation enthalpy ( $\Delta H$ ). Negative formation enthalpy means an exothermic process, and the lower formation energy indicates the stability with respect to the decomposition to elemental constituents. The formation enthalpy can be calculated by the following expression [41]:

$$\Delta H = (E_{\text{tot}} - \sum n_i E_i) / n \quad (11)$$

where  $E_{\text{tot}}$  is the total energy of the bulk compound with  $n_i$  atoms of all  $i$  (Tb and P),  $n$  is the total number of atoms in the primitive cell, and  $E_i$  is the total energy of a pure  $i$  atom with equilibrium lattice parameters. The calculated theoretical formation enthalpies of TbP compounds are included in Table 1. Unfortunately as far as it is known, there are no data available related to formation energy in the literature to compare with its. Negative formation enthalpies indicate their structural stabilities from energetic point of view. B8 structure of TbP shows the lowest value of formation enthalpy, which indicates that B8 structure of TbP has the highest structural stability which is compatible with Figure 1.

Our computational approach is based on constant-pressure static quantum mechanical calculations at  $T=0$  K, so the relative stability of different phases can be deduced from the pressure dependence of the enthalpy instead of the Gibbs free energy [42]. The pressure-volume curve is plotted for both B1 and B2 structures of TbP in Figure 2. Naturally, the cell volume decreases with increasing pressure values. The

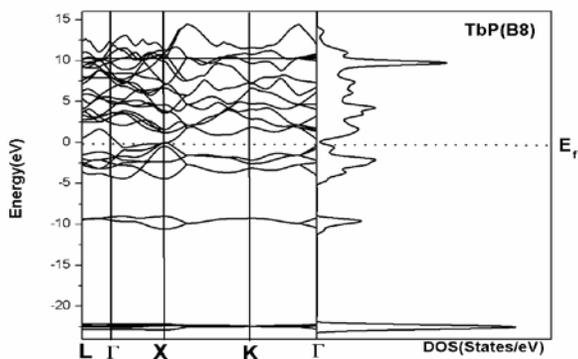
discontinuity in volume takes place at the phase transition pressure. The phase transition pressures from B1 to B2 structure are found to be 55 GPa TbP.



**Fig. 3.** Enthalpy-pressure curves for B1 and B2 structures of TbP

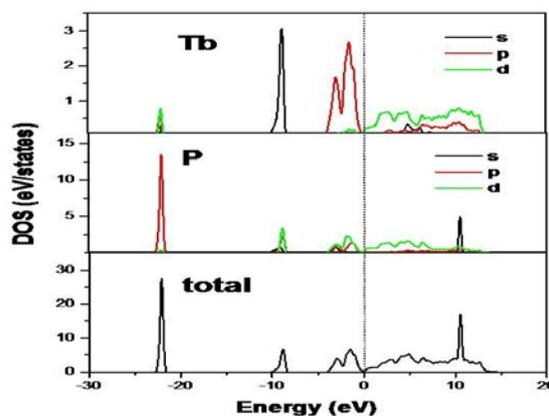
The related enthalpy versus pressure graph is shown in Figure 3 for TbP. The transition pressure is a pressure at which H(p) curves for both structures cross. The same result is also confirmed in terms of the common tangent technique in Figure 1.

In order to understand the electronic and phase stability of TbP the energy band structure along with total electronic density of states at 0 GPa for B8 are presented in Figure 4. Fermi level is set 0eV. Our calculation shows that the B8 structure of TbP is of metallic conductivity as there is no band gap near the Fermi level and there are many bands crossing the Fermi level.

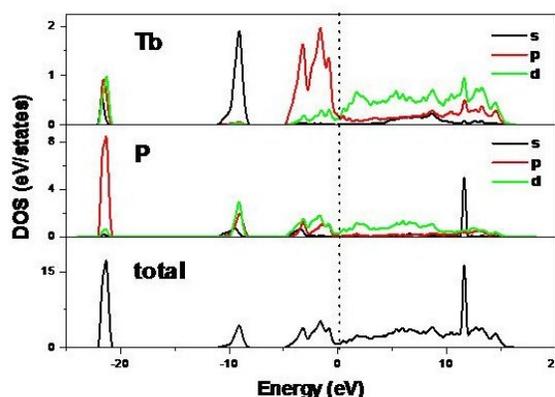


**Fig. 4.** Electronic band structure and total density of states of TbP(B8).

(a)



(b)



**Fig. 5.** Partial and total DOS of TbP(B8) at (a) 0 GPa (b)50GPa

Partial DOS figures are also presented at 0 GPa and 50 GPa for TbP in B8 structure in Figure 5. The lowest valance bands occur between about -10.5 and 5.5 eV are essentially dominated by Tb-s states. Other valance bands are essentially dominated by Tb-p, Tb-d, P-p and P-d states. In Fig. 5 (b), the lowest valance bands occur between about -11 and 12 eV are dominated by Tb-s states. Tb-d and P-d states are more robust at 50 GPa according to 0 GPa. Contribution of the P-p state at 0 GPa is more than at 50 GPa.

#### Elastic properties

Elastic properties of materials are very important because of various fundamental solid state properties, such as Zener anisotropy factor, shear modulus, Poisson's ratio, Young modulus and so on. The elastic constants determine the response of the crystal to external forces and play a big role in determining the strength of the materials.

There are two common methods [43, 44] for obtaining the elastic constants through the ab-initio modelling of materials from their known crystal structures: an approach based on analysis of the total energy of properly strained states of the material in the volume conserving technique and an approach based on the analysis of changes in calculated stress values resulting from changes in the stress-strain technique. In this

work, the stress-strain technique is used for obtaining the second-order elastic constants ( $C_{ij}$ ). The stress-strain technique is based on constructing a set of linear equations from stress-strain relation for several deformations of the unit cell. This set of equations represents a general form of Hook's law and can be solved with respect to the elastic constants.

$$\sigma_i = \sum_{j=1}^6 C_{ij} \varepsilon_j, \quad (11)$$

that describes the linear dependency of stress component  $\sigma_i$  ( $i = 1-6$ ) and applied strain  $\varepsilon_j$  ( $j = 1-6$ ) under a small deformation. Here  $C_{ij}$  are the elastic constants of the crystal whose structure has been fully relaxed under a given set of

exchange-correlation potential functions and attained an equilibrium structure with a minimum total energy. In order to obtain the elastic constants, we calculate the second derivatives of the internal energy with respect to the strain tensor. The Born's stability criteria's [45] should be satisfied for the stability of lattice. The known conditions for mechanical stability of cubic crystals are:  $C_{11}>0$ ,  $C_{11}-C_{12}>0$ ,  $C_{44}>0$ ,  $C_{11}+2C_{12}>0$  and  $C_{12}<B<C_{11}$ . For hexagonal structure the mechanical stability criteria are given by  $C_{44}>0$ ,  $C_{11}>C_{12}$ ,  $(C_{11} + 2C_{12})C_{33} > 2C_{13}^2$ . For tetragonal structure the mechanical stability criteria are  $C_{11}>0$ ,  $C_{33}>0$ ,  $C_{44}>0$ ,  $C_{66}>0$ ,  $(C_{11}-C_{12})>0$ ,  $(C_{11}+C_{33}-2C_{12})>0$  and  $2(C_{11}+C_{12})+C_{33}+4C_{13}>0$ .

**Table 2.** The calculated elastic constants (in GPa unit) in different structures for TbP.

| Structure       | Reference | $C_{11}$ | $C_{12}$ | $C_{44}$ | $C_{33}$ | $C_{13}$ | $C_{66}$ | Stability       |
|-----------------|-----------|----------|----------|----------|----------|----------|----------|-----------------|
| B1              | Present   | 211.06   | 26.81    | 45.56    |          |          |          | <i>Stable</i>   |
| B2              | Present   | 151.16   | 55.15    | 38.63    |          |          |          | <i>Stable</i>   |
| B3              | Present   | 60.54    | 54.01    | 39.78    |          |          |          | <i>Stable</i>   |
| Bh              | Present   | 139.56   | 51.34    | 44.11    | 207.57   | 35.38    | 19.00    | <i>Stable</i>   |
| L1 <sub>0</sub> | Present   | 13.89    | 166.98   | 41.64    | 173.20   | 62.91    | 39.09    | <i>Unstable</i> |
| B8              | Present   | 147.14   | 44.95    | 51.10    | 197.81   | 40.51    | 56.66    | <i>Stable</i>   |
| B10             | Present   | 180.30   | 64.34    | 52.77    | 42.80    | 23.61    | 6.69     | <i>Stable</i>   |
| B4              | Present   | 97.99    | 41.63    | 28.19    | 114.65   | 29.74    | 20.40    | <i>Stable</i>   |

The calculated values of  $C_{ij}$  are given in Table 2 for TbP compound. The related mechanical stability conditions are satisfied except for L1<sub>0</sub> structure in TbP. The L1<sub>0</sub> structure in TbP is mechanically unstable despite the fact that all other structures in TbP are mechanically stable. For cubic structures (B1, B2 and B3)  $C_{11}$  are higher than  $C_{12}$  and  $C_{44}$  and other structures, the values of  $C_{11}$  and  $C_{33}$  are much higher than those of  $C_{12}$ ;  $C_{13}$ ;  $C_{44}$  and  $C_{66}$ , indicating TbP compound under investigation is mechanically anisotropic and the shear deformation is easier to take place than compression deformations along the principle direction a- and c-axis. To our best knowledge, no experimental and theoretical data are available in the literature to be compared with our results. Then, our results can serve as a prediction for future investigations.

The Zener anisotropy factor  $A$ , Poisson's ratio  $\nu$ , and Young's modulus  $Y$ , which are the most interesting elastic

properties for applications, are also calculated in terms of the computed using the following relations [46]:

$$A = \frac{2C_{44}}{C_{11} - C_{12}}, \quad (12)$$

$$\nu = \frac{1}{2} \left[ \frac{(B - \frac{2}{3}G)}{(B + \frac{1}{3}G)} \right], \quad (13)$$

$$Y = \frac{9GB}{G + 3B} \quad (14)$$

where  $G=(G_V+G_R)/2$  is the isotropic shear modulus,  $G_V$  is Voigt's shear modulus corresponding to the upper bound of  $G$  values and  $G_R$  is Reuss's shear modulus corresponding to the

lower bound of G values and can be written as  $G_V=(C_{11}-C_{12}+3C_{44})/5$  and  $5/G_R= 4/(C_{11}-C_{12})+3/C_{44}$ . The calculated Zener anisotropy factor (A), Poisson's ratio ( $\nu$ ), Young's modulus (Y) and Shear modulus ( $C'=(C_{11}-C_{12}+2C_{44})/4$ ) for TbP are given in Table 3 and they are close to these obtained for the similar structural symmetries.

**Table 3.** The calculated Zener anisotropy factor (A), Poisson's ratio ( $\nu$ ), Young's modulus (Y), shear modulus ( $C'$ ) for TbP in B8 phase.

| Material | A | $\nu$  | Y<br>(GPa) | $C'$<br>(GPa) | B/G  |
|----------|---|--------|------------|---------------|------|
| TbP      | 1 | 0.2070 | 137.17     | 51.10         | 1.53 |

The Zener anisotropy factor (A) takes the value of 1 for a completely isotropic material that shows this compound is completely isotropic materials. The Poisson's ratio ( $\nu$ ) characterizes the stability of the crystal against shearing strain. For a typical metal, the value is supposed to be 0.33; for the ionic-covalent crystal, the value is to be between 0.2 and 0.3. We obtain 0.207 that is situated ionic-covalent crystal. The Young modulus (Y) which is calculated 137.17 GPa is measurement of the stiffness of the solids. B/G ratios that are roughly as a measurement of brittleness or ductility are also given in Table 3. Providing that the critical value is 1.75 and/or more than this value, the material is regarded as ductile. [47-49]. TbP in B8 structure indicates brittle behavior due to the fact that the present value of B/G is 1.53. As seen from table 3, the values of B/G ratio for TbP compound is smaller than the critical value, thus this compounds in B8 structure may be classified as brittle material. Mechanical properties such as ductility and brittleness of semiconductor materials are very important for their technological applications.

#### Thermodynamic properties

Thermodynamic properties are determined in the temperature range 0-2000 K and the pressure range 0-50 GPa for B8 structure of TbP. The calculations based on the first principles methods demonstrate that quasi-harmonic approximation provides a reasonable description of the dynamic properties of many bulk materials below the melting point [50-54]. The melting point of TbP is calculated to be  $1422\pm 300$  K. Hence, for decreasing the probable influence of anharmonicity, where the quasi-harmonic model remains fully valid. The Debye temperature ( $\theta_D$ ) is known as an important fundamental parameter closely related to many physical properties such as specific heat and melting temperature. At low temperatures the vibrational excitations arise solely from acoustic vibrations. Hence, at low temperatures the Debye temperature calculated from elastic constants is the same as that determined from specific heat measurements. We have calculated the Debye temperature,  $\theta_D$ , from the elastic constants using the average sound velocity,  $v_m$ , by the following common relation given [55]

$$\theta_D = \frac{\hbar}{k} \left[ \frac{3n}{4\pi} \left( \frac{N_A \rho}{M} \right) \right]^{1/3} v_m \quad (15)$$

where  $\hbar$  is Planck's constants,  $k$  is Boltzmann's constants,  $N_A$  Avogadro's number,  $n$  is the number of atoms per formula unit,  $M$  is the molecular mass per formula unit,  $\rho(=M/V)$  is the density, and  $v_m$  is obtained from

$$v_m = \left[ \frac{1}{3} \left( \frac{2}{v_l^3} + \frac{1}{v_t^3} \right) \right]^{-1/3} \quad (16)$$

where  $v_l$  and  $v_t$ , are the longitudinal and transverse elastic wave velocities, respectively, which are obtained from Navier's equations [56]:

$$v_l = \sqrt{\frac{3B+4G}{3\rho}} \quad (17)$$

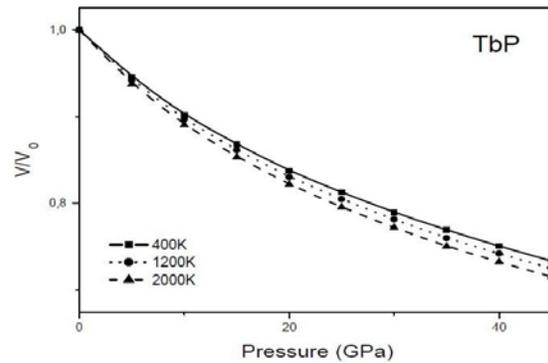
and

$$v_t = \sqrt{\frac{G}{\rho}} \quad (18)$$

**Table 4.** The longitudinal ( $v_l$ ), transverse ( $v_t$ ), average ( $v_m$ ) elastic wave velocities, Debye temperature ( $\theta_D$ ) and melting temperature ( $T_m$ ) for TbP(B8).

| Material | $v_l$ (m/s) | $v_t$ (m/s) | $v_m$ (m/s) | $\theta_D$ (K) | $T_m$ (K)        |
|----------|-------------|-------------|-------------|----------------|------------------|
| TbP      | 6873.4      | 4177.8      | 4615.7      | 377.0          | $1422.6 \pm 300$ |

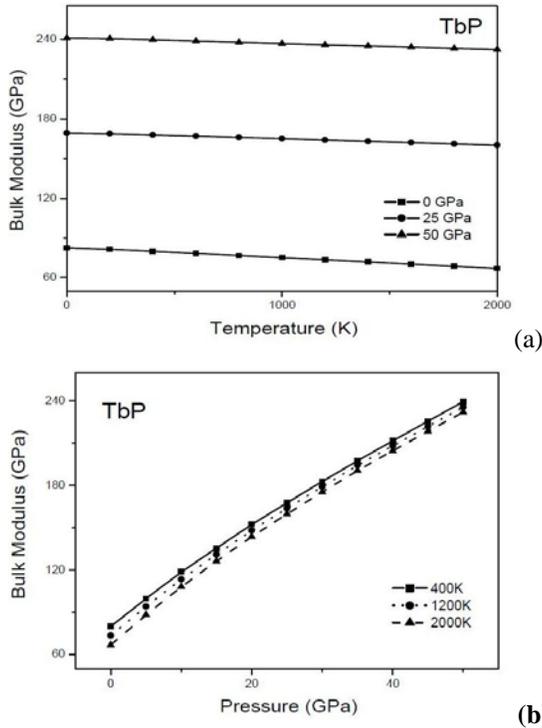
The calculated longitudinal, transverse and average elastic wave velocities, Debye temperature and melting temperature for TbP are given in Table 4. No other theoretical or experimental values are exist for comparison with these present values.



**Fig. 6.** The normalized volume-pressure curves for B8 structure of TbP at different temperatures

The relationship between normalized volume and pressure at different temperature is shown in Figure 6 for TbP. It can

be seen that when the pressure increases from 0 GPa to 50 GPa the volume decreases. The reason of this changing can be attributed to the atoms in the interlayer become closer, and their interactions become stronger.



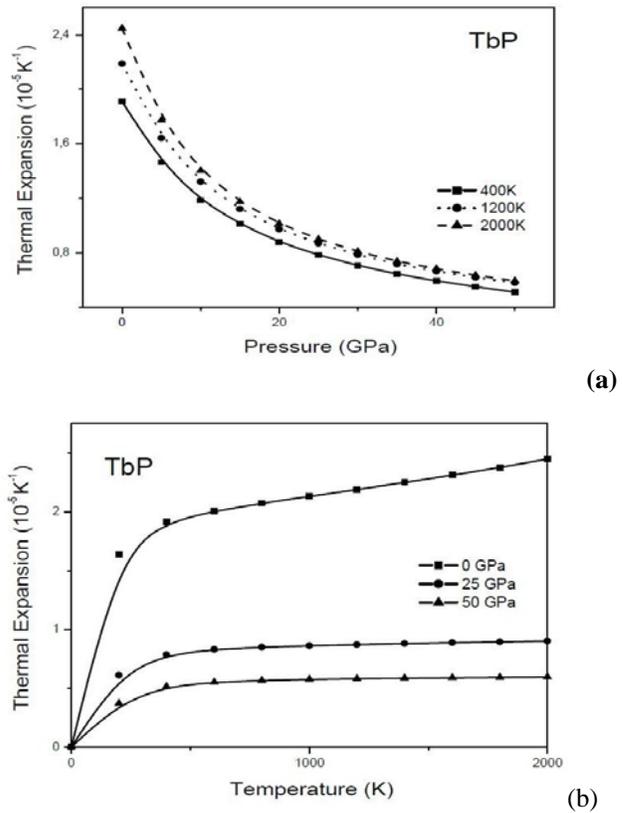
**Figure 7. (a)** The variations of bulk modulus with temperature for B8 structure of TbP.

**(b)** The variations of bulk modulus with pressure for B8 structure of TbP.

The variations of bulk modulus, B, with temperatures and pressures are presented in Fig. 7 for B8 structure. It can be easily seen that B decreases as temperature increases in Figure 7 (a). Because cell volume changes rapidly as temperature increases. Relations of bulk modulus and temperature polynomial curves are third-order fitted and performed for stable structure TbP(B8). Relation is given as below at 0 GPa for TbP.

$$B = 82.65455 - 6.06 \times 10^{-3} T - 1.75699 \times 10^{-6} T^2 + 4.23951 \times 10^{-10} T^3$$

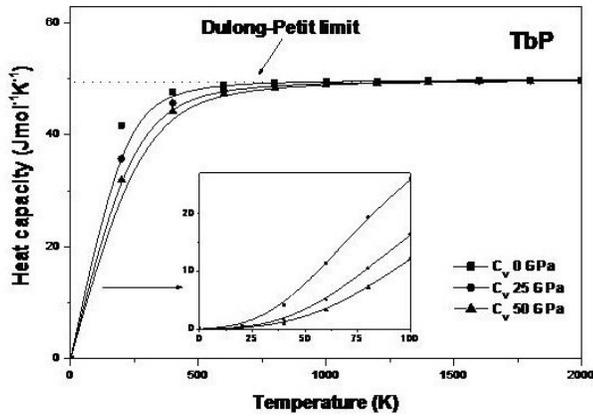
The relationship between bulk modulus and pressure at different temperatures (400, 1200 and 2000K) is shown in Figure 7(b) for TbP. It can be seen that bulk modulus decreases with the temperature at a given pressure and increases with pressure at a given temperature. It shows that the effect of increasing pressure on TbP is the same as decreasing its temperature.



**Figure 8.(a)** The thermal expansion coefficient versus pressure at different temperatures for TbP in B8 structure.

**(b)** The thermal expansion coefficient versus temperature at different pressures for TbP in B8 structure.

The variations of the thermal expansion coefficient ( $\alpha$ ) with pressure and temperature are shown in Figure 8 for TbP. At different temperatures, thermal expansion coefficient decreases nonlinearly at lower pressure and decreases almost linearly at higher pressure values in Fig. 8(a). The thermal expansion coefficient increases at lower temperatures and gradually approaches linear increases at higher temperatures, while the thermal expansion coefficient decreases with pressure in Fig. 8(b).

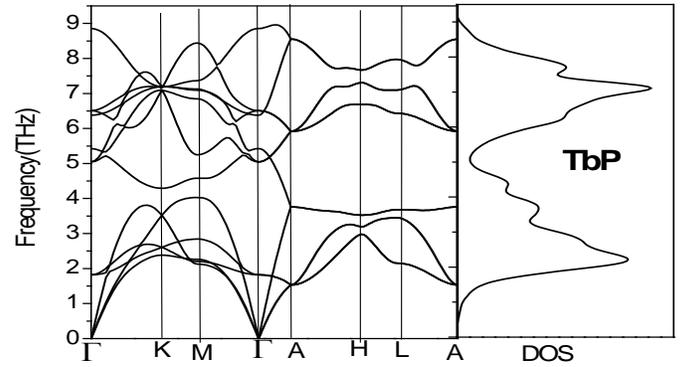


**Fig. 9.** The variation of  $C_v$  with temperature at different pressures for TbP.

The variations of the temperature dependence of the heat capacity at constant volume,  $C_v$ , at different pressures, presented in Figure 9 for TbP follows the Debye Law that is at low temperatures ( $T < 250$  K),  $C_v$  is proportional to  $T^3$  and high temperatures ( $T > 500$  K) the heat capacity  $C_v$  is very close to Dulong-Petit limit.

#### Phonon Dispersion Curves

The phonon dispersion curves and phonon density of states for B8 structure of TbP are calculated by using the PHONOPY program [57] using the interatomic force constants obtained from VASP [23-25] which is use linear response method within the density functional perturbation theory (DFPT) [58- 60]. The Phonopy program which is based on real space supercell calculates phonon frequencies from force constants. The obtained phonon dispersion curves and the corresponding one-phonon DOS for TbP along the high symmetry directions using a  $2 \times 2 \times 2$  cubic supercell of 32 atoms are illustrated in Fig. 10. The absence of the soft modes in the phonon dispersion curves confirms the dynamical stability of TbP. To our knowledge there are no experimental and other theoretical works exploring the lattice dynamics of this compound for comparison with the present data; hence our work is a first attempt in this direction. Owing to the mass difference between Tb and P atom the no band gap takes place between acoustic and optical regions. On the right side of phonon curves in Fig. 10, the corresponding total density of phonon states for this compound is shown.



**Fig. 10.** The calculated phonon dispersions and density of states for TbP in B8 structure.

#### IV. SUMMARY AND CONCLUSION

In our detailed work, we have investigated the structural, elastic, electronic and thermodynamic and vibrational properties properties of TbP in different structures using ab-initio plane wave pseudopotentials density functional theory (DFT). The lattice parameters are calculated and found good agreement with other experimental values for B1 structure. On the basis of using common tangent method, we have found that phase transition from B1 to B2 phase occurs at about 55 GPa for TbP. Electronic calculations are presented with band structures, total density of states and partial DOS. The electronic properties are performed both at 0 GPa and 50 GPa pressure values. The zero pressure second order elastic constants and their related quantities such as bulk modulus, shear modulus, Young's modulus, Zener anisotropy factor, Poisson ratio and B/G ratios are investigated and decided to brittle/ductile behavior for this compound which shows brittle behavior. The quasi-harmonic Debye model is successfully used for thermodynamic properties calculations in the wide range temperature and pressure values. The longitudinal, transverse, average elastic wave velocities, Debye temperatures and melting points for TbP in B8 structure are calculated and given in related tables. The absence of the soft modes in the phonon dispersion curves confirms the dynamical stability of TbP. To the best of our knowledge, our calculations are the first theoretical prediction on the thermodynamic

#### ACKNOWLEDGMENT

Acknowledgments Authors want to express their great acknowledges to the scientific research unit of Amasya University for the financial support to this study with Grant Number of FMB-BAP-13-059.

#### REFERENCES

- [1] A. Buschbeck, G.H. Chojnowski, J. Kötzler, R. Sonder and G. Thummes, "Field-dependent phase transitions and magnetization of the type II- antiferromagnets TbP and TbSb", Journal of Magnetism and Magnetic Materials, vol. 69, Oct. 1987, pp. 171-182.

- [2] L. Petit, R. Tyer, Z. Szotek, W.M. Temmerman and A. Svane, "Rare earth mononictides and monochalcogenides from first principles: towards an electronic phase diagram of strongly correlated materials" *New J. Phys.* vol. 12, Nov. 2010 113041 (20pp).
- [3] G. Pagare, S.S. Chouhan, P. Soni, S.P. Sanyal, M. Rajagopalan, "First principles study of structural, electronic and elastic properties of lutetium mono-pnictides" *Comp. Mater. Sci.* vol. 50, Dec. 2010 538-544.
- [4] G. Bruzzone, A.F. Ruggiero, "The equilibrium diagram of the calcium-indium system" *J. Less-Common Met.* Vol. 7 Nov. 1964, 368-372.
- [5] G. Bruzzone, "Sui sistemi binari Sr-Tl, Ba-Tl e Ca-Tl." *Annali di Chimica, Rome* vol. 56 1966, 1306-1319.
- [6] J.D. Marcoll, P.C. Schmidt, A. Weiss, Z. Naturforsch, "X-Ray investigations of intermetallic phases CaCd<sub>1</sub>-XTiX and CaIn<sub>1</sub>-XTiX and knight-shift measurements of Tl-205-NMR and Cd-113-NMR in system CaCd<sub>1</sub>-XTiX." *Zeitschrift Fur Naturforschung Section AA J. Phys. Scienc.* vol. 3, 1974, 473-476.
- [7] J. Rosat-Mignod, J.M. Effantin, P. Bulet, T. Chattopadhyay, L.P. Regnault, H. Bartholin, C. Vettier, "O. Vogt, D. Ravot and JC Achard.", *J. Magn. Magn. Mater.* vol. 52, 1985, 111.
- [8] T. Chattopadhyay, P. Bulet, J. Rosat-Mignod, H. Bartholin, C. Vettier, and O. Vogt, "High-pressure neutron and magnetization investigations of the magnetic ordering in CeSb" *Phys. Rev. B* vol. 49, Jun. 1994, 15096.
- [9] R. Pittini, J. Schoenes, O. Vogt, and P. Wachter, "Discovery of 90 degree Magneto-optical Polar Kerr Rotation in CeSb" *Phys. Rev. Lett.* vol.77, Jul. 1996, 944.
- [10] R. Pittini, J. Schoenes, F. Hulliger, and P. Wachter, "Periodicity of the Spin Structure Observed in the Optical Response of CeBi Single Crystals" *Phys. Rev. Lett.* vol. 76, Apr. 1996, 3428.
- [11] O. Vogt and K. Mattenberger, "The extraordinary case of CeSb" *Physica B: Cond. Matt.* vol. 215, Oct. 1995, 22-26.
- [12] J. Schoenes, P. Repond, F. Hulliger, D.B. Ghosh, S.K. De, J. Kunes, P.M. Oppeneer, "Experimental and theoretical investigation of optical properties of dysprosium mononictides" *Phys. Rev. B* vol. 68, Aug. 2003, 085102.
- [13] E. Zintl, C. Brauer, *Z. Phys. Chem., Abt. B* vol. 20 1933, 245-271.
- [14] Chun-Gang Duan, R.F. Sabirianov, W.N. Mei, P.A. Dowben, S.S. Jaswall and E.Y. Tsymbal, "Electronic, magnetic and transport properties of rare-earth mononictides" *Journal of Physics: Condensed Matter* vol.19, Jul. 2007, 315220.
- [15] E. Zintl, S. Neumayr, "Gitterstruktur des Indiums. (7. Mitteilung über Metalle und Legierungen)" *Z. Elektrochem.* vol. 39, Febr. 1933, 81-84.
- [16] B.D. Cullity, *Introduction to Magnetic Materials*, in: M. Cohen Ed., Addison-Wesley Publishing, Reading, Mas-Ž. sachusetts, 1972, p. 178.
- [17] A. Hasegawa and A. Yanase, "Energy Band Structures of Gd-Pnictides" *J. Phys. Soc. Japan*, vol. 42, 1977, 492-498.
- [18] Y. Nakanishi, T. Sakon, M. Motokawa, T. Suzuki, "De Haas-van Alphen study of the spin splitting of the Fermi surface in TbSb" *Phys. Rev. B* vol. 69, Jun. 2004, 024412-6.
- [19] Y. Nakanishi, T. Sakon, M. Motokawa, T. Suzuki, M. Yoshizawa, "Elastic properties and phase diagram of the rare-earth mononictide TbSb" *Phys. Rev. B* vol. 69, Oct. 2003, 144427-6.
- [20] S.P. Gordienko, "Reaction of Titanium with Boron Nitride under Self-Propagating High-Temperature Synthesis Conditions" *Powder Metallurgy and Metal Ceramics*, vol. 40, Jan. 2001, 58-60.
- [21] Y.O. Ciftci, M. Ozayman, G. Surucu, K. Colakoglu, E. Deligoz, "Structural, electronic, elastic, thermodynamic and vibration properties of TbN compound from first principles calculations" *Solid State Sciences* vol.14, Mar. 2012, 401-408.
- [22] J. Kötzer, G. Raffius, A. Loidl and C.M.E. Zeyen, "Singlet-groundstate magnetism in TbP" *Z. Physik, B* vol. 35, May. 1979, 125-132.
- [23] G. Kresse and J. Hafner, "Ab. initio molecular dynamics for liquid metals" *Phys. Rev. B* vol. 47, Jan. 1993, 558-561.
- [24] G. Kresse and J. Furthmauller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set" *Phys. Rev. B* vol. 54, Oct. 1996, 11169-11186.
- [25] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method" *Phys. Rev. B* vopl. 59, Jan. 1999, 1758-1775.
- [26] P.E. Blochl, "Projector augmented-wave method" *Phys. Rev. B* vol. 50, Dec. 1994, 17953-17979.
- [27] J.P. Perdew and A. Zunger, "Self-interaction correction to density-functional approximations for many-electron systems" *Phys. Rev. B* vol. 23, May. 1981, 5048-5079.
- [28] J.P. Perdew, J.A. Chevary, S.H. Vosko, K.A. Jackson, M.R. Pederson, D.J. Singh and C. Fiolhais, "Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation" *Phys. Rev. B* vol. 46, Sep. 1992, 6671-6687.
- [29] H.J. Monkhorst and J.D. Pack, "Special points for Brillouin-zone integrations" *Phys. Rev. B*, vol. 13, Jun. 1976 5188-5192.
- [30] M.A. Blanco, E. Francisco, V. Lunana, "GIBBS: isothermal-isobaric thermodynamics of solids from energy curves using a quasi-harmonic Debye model" *Comput. Phys. Commun.* vol. 158, Marc. 2004, 57-72.
- [31] M.A. Blanco, A.M. Pendàs, E. Francisco, J.M. Recio, R. Franko, "Thermodynamical properties of solids from microscopic theory: applications to MgF<sub>2</sub> and Al<sub>2</sub>O<sub>3</sub>" *J.Mol. Struct. (Theochem)* vol. 368, Sep. 1996, 245-255.
- [32] M. Flórez, J.M. Recio, E. Francisco, M.A. Blanco, A.M. Pendàs, "First-principles study of the rocksalt-cesium chloride relative phase stability in alkali halides" *Phys. Rev. B* vol. 66, Oct. 2002, 144112-8.
- [33] E. Francisco, J.M. Recio, M.A. Blanco, A.M. Pendàs, "Quantum-Mechanical Study of Thermodynamic and Bonding Properties of MgF<sub>2</sub>" *J. Phys. Chem.* vol.102, (1998) 1595-1601.
- [34] R. Hill, "The Elastic Behaviour of a Crystalline Aggregate" *Proc. Phys. Soc. Lond. A* vol. 65, 1952, 349-354.
- [35] F.D. Murnaghan, *Proc. Natl., Acad. Sci. USA* vol. 30, 1944, 5390.
- [36] F. Lévy F, *Phys. Kondens. Mater.* vol. 10, 1969, 85-106.
- [37] K. Yaguchi, *J. Phys. Soc. Jpn.* vol. 21, 1996, 1226.
- [38] H.R Child, M.K. Wilkinson, J.W. Cable, W.C. Koehler, E.O. Wollan, "Neutron Diffraction Investigation of the Magnetic Properties of Compounds of Rare-Earth Metals with Grouy V Anions" *Phys. Rev.* vol. 131, Aug. 1963, 922-931.
- [39] M.K. Wilkinson, H.R Child, W.C. Koehler, J.W. Cable, E.O. Wollan, "Recent Magnetic Neutron Scattering Investigations at Oak Ridge National Laboratory" *J. Phys. Soc. Jpn.* vol. 17, 1962, 27-31.
- [40] G.L. Olcese, G.B. Bonino, *Atti Accad. Naz. Lincei, Cl. Sci. Fis., Mat. Nat., Rend.* vol. 30 1961, 195-200.
- [41] Y.P. Xie, Z.Y. Wang, Z.F. Hou, "The phase stability and elastic properties of MgZn<sub>2</sub> and Mg<sub>4</sub>Zn<sub>7</sub> in Mg-Zn alloys" *Scr. Mater.* vol. 68, Apr. 2013, 495-498.
- [42] A. Hao, Y. Zhu, "First-principle investigations of structural stability of beryllium under high pressure" *J. Appl. Phys.* vol.112, Jul. 2012, 023519-4.
- [43] J. Mehl, "Pressure dependence of the elastic moduli in aluminum-rich Al-Li compounds" *Phys. Rev. B* vol. 47, Feb. 1993, 2493-2500.
- [44] S.Q. Wang, H.Q. Ye, "First-principles study on elastic properties and phase stability of III-V compounds" *Phys. Status Solidi B* vol. 240, 2003, 45-54.
- [45] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, Clarendon, Oxford, 1956.
- [46] B. Mayer, H. Anton, E. Bott, M. Methfessel, J. Sticht, and P. C. Schmidt, "Ab-initio calculation of the elastic constants and thermal expansion coefficients of Laves phases" *Intermetallics* vol. 11, Jan. 2003, 23-32.
- [47] S.F. Pugh, "XCII. Relations between the elastic moduli and the plastic properties of polycrystalline pure metals" *Phil. Mag.* vol. 45, Apr. 1954, 823-843.
- [48] V.V. Bannikov, I.R. Shein, A.L. "Electronic structure, chemical bonding and elastic properties of the first thorium-containing nitride perovskite TaThN<sub>3</sub>" *Ivanovskii Phys Status Solidi (RRL)* vol. 1, May. 2007, 89-91.
- [49] I. Johnston, G.Keeler, R. Rollins, and S.Spicklemyre, *Solid State Physics Simulations, The Consortium for Upper-Level Physics Software*, John Wiley, New York, 1996.
- [50] S. Biernacki, M. Scheffler, "Negative Thermal Expansion of Diamond and Zinc-Blende Semiconductors" *Phys. Rev. Lett.* vol.63 Jul. 1989, 290-293.
- [51] A. Fleszar, X. Gonze, "First-principles thermodynamical properties of semiconductors" *Phys. Rev. Lett.* vol. 64, Jun. 1990, 2961.
- [52] P. Pavone, K. Karch, O. Schutt, W. Windl, D. Strauch, P. Giannozzi, S. Baroni, "Ab initio lattice dynamics of diamond" *Phys. Rev. B* vol. 48, Aug. 1993, 3156-3163.
- [53] P. Pavone, S. Baroni, S. De Gironcoli, "a↔b phase transition in tin: A theoretical study based on density-functional perturbation theory" *Phys. Rev. B* vol. 48, May. 1998, 10421-10423.

- [54] A.A. Quang, A.Y. Liu, "First-principles calculations of the thermal expansion of metals" *Phys. Rev. B* vol. 56, Oct. 1997, 7767-7770.
- [55] E. Schreiber, O. L. Anderson, N. Soga, *Elastic Constants and Their Measurements*, McGraw-Hill, New York, 1973.
- [56] M.E. Fine, L.D. Brown, H.L. Marcus, *Scr. Metall.* vol.18, 1984, 951.
- [57] A. Togo, F. Oba, and I. Tanaka, *Phys. Rev. B*, vol. 78, 2008, 134106-1-9.
- [58] S. Baroni, P. Giannozzi, and A. Testa "Green's-function approach to linear response in solids" *Phys. Rev. Lett.*, vol. 58, May. 1987, 1861-1864.
- [59] X. Gonze and J.-P. Vigneron, "Density-functional approach to nonlinear-response coefficients of solids" *Phys. Rev. B*, vol. 39, Jun. 1989, 13120-13128.
- [60] X. Gonze, D. C. Allan, and M. P. Teter "Dielectric Tensor, Effective Charges, and Phonons in  $\alpha$ -Quartz by Variational Density-Functional Perturbation Theory" *Phys. Rev. Lett.*, vol. 68 Jun. 1992, 3603-3606.

## Decision making in Group process of consensus based on structures of Decision Dynamics: application to the Superior Council of the UTEM

MUÑOZ S. SIMÓN, ZAPATA C. SANTIAGO

Department of Informatics and computing

Faculty of engineering, Technological Metropolitan University of the State of Chile (UTEM)

José Pedro Alessandri 1242, Macul

SANTIAGO, CHILE, CP 8330378

simon.munoz.saavedra@gmail.com, szapata@utem.cl, <http://www.utem.cl>

### Abstract

Group Decision making is a complex process due to the participation of many experts, which issued their opinions regarding a problem with different alternatives in order to give a solution in common agreement. To combat this complexity sets a process of consensus, where experts align their positions to reach a solution. With this, the consensus process be held in rounds, in which each expert will have the opportunity to modify their preferences together and thus align their views. It may be the case in which one or more experts do not engage in any of the rounds of decision, or also, new experts join the rounds. In these cases, it is necessary to establish a mechanism that allows remember the opinions expressed by the experts at any moment in time, so it makes use of dynamic decision structures which allow you to store information about the rounds of decision of the consensus process. This contribution presents a practical example of the application of this model of decision making in group based on the sessions of the Superior Council of the UTEM directory.

## 1 Introduction

Decision making is an action that takes place in everyday life. It is present in events as simple as the choice of what will be taken from breakfast, to the choice of a career. Throughout the day people are faced with a world of choices, these consist of multiple alternatives of which shall be selected which are more adapted to the needs or objectives that have.

On many occasions take a decision may seem quite simple, but the decision-making process is somewhat complex and requires a prior analysis of the various alternatives presented before a problem. Here you have to evaluate the advantages and disadvantages of each alternative and the impact that these would cause in the outcome

of the election. In addition to the above, there are problems of decision making according to different points of view, which can go from the number of people who participate in a decision, to the context in which develops the decision. When involving more than one person, or expert, in a decision problem, it is called Group Decision Making (GDM).

The Group decision making is a process even more complex since it involves more people, therefore, the more points of view, more views, different criteria, etc. , and the idea, is to ensure that all these people are in agreement, which in practice is an almost impossible task, for which we are talking about reaching an agreement where the great majority is in favor of this.

## 2 The Group Decision Making

The group decision making [27] is the process in which a group of experts is trying to solve a problem through the selection of different alternatives from a set of these associated with this problem, which constitute the solution of the same, this solution should be common to all the experts.

When you have a problem, and this is resolved by a group of people by applying different criteria and points of view, there is a very high probability that the solution to this problem is of higher quality than the solution provided by a unique individual, since the latter has a thought reduced in comparison to the collective thinking.

In a formal manner, a process of group decision making is characterized by [13]:

- The existence of a problem that want to be solved.
- A set denoted by  $X = \{x_1, \dots, x_n\}$  with  $(n \geq 2)$ , of possible alternatives or solutions to the problem raised.
- A set denoted by  $E = \{e_1, \dots, e_m\}$  with  $(m \geq 2)$ ,

or makers of experts which provide their assessments or preferences on the set of alternatives to the problem raised.

The contribution of valuations that the experts assigned to the alternatives ready for the problem that arises is done through structures of preference. As we saw above, it is called **preference** to a favorable attitude expressed by the decision maker to any alternative, once compared to another. Currently, the structure of preference more used in the problems of group decision making under uncertainty is the *Fuzzy preference relation* [7, 13, 14]. This structure makes use of a diffuse criterion for the assignment of a value toward an alternative, where a relationship of fuzzy preference denoted by  $P_i$  of the expert  $e_i$  is characterized by a preference feature:  $\mu_{p_i} : X \times X \rightarrow [0, 1]$ , where its representation on the set of alternatives is an array of dimension  $n \times n$ :

$$P_i = \begin{pmatrix} p_i^{11} & \dots & p_i^{1n} \\ \vdots & \ddots & \vdots \\ p_i^{n1} & \dots & p_i^{nn} \end{pmatrix}$$

Where each valuation  $p_i^{lk} = \mu_{p_i}(x_l, x_k)$  represents the degree of preference of the alternative  $x_l$  on  $x_k$  according to the expert  $e_i$ , so that  $p_i^{lk} > 0.5$  indicates the preference of  $x_l$  on  $x_k$ ,  $p_i^{lk} < 0.5$  indicates the preference of  $x_k$  on  $x_l$ , and  $p_i^{lk} = 0.5$  indicates the indifference between  $x_l$  and  $x_k$ .

An example of the use of the array of preference for a particular problem that has five alternatives could be:

$$P_1 = \begin{pmatrix} 0.500 & 0.439 & 0.373 & 0.556 & 0.649 \\ 0.561 & 0.500 & 0.583 & 0.651 & 0.615 \\ 0.627 & 0.417 & 0.500 & 0.709 & 0.680 \\ 0.444 & 0.349 & 0.291 & 0.500 & 0.603 \\ 0.351 & 0.385 & 0.320 & 0.397 & 0.500 \end{pmatrix}$$

In addition, as you can see, the array of preference relations account with the following properties:

- The preference of an alternative to itself is: 0.5.
- Reciprocity, where:  $p_i^{lk} + p_i^{k1} = 1$ .

In general the selection of alternatives in problems of group decision making is composed of two phases:

1. **Aggregation Phase:** This phase consists in transforming a set of elements (diffuse, individual opinions on a set of alternatives, etc.) in a single representative of the same item. In Group decision making problems the aggregation phase consists of the combination of individual information units in collective information units.

2. **Exploitation Phase:** This is the last step in the process of group decision making, this phase uses the information provided by the phase of aggregation to identify the whole solution to the problem. This process seeks to transform the global information on the alternatives in a holistic management of the same. It must have been previously set a selection criterion to establish an order among the set of alternatives of the problem.

## 2.1 Consensus Process

While the group decision making [27] is a collective process there may be differences in the selection process of alternatives in decision problems with multiple experts, which can lead to solutions that are not accepted as good by the whole group, so that the study of the consensus has become a field of research of great importance within the decision-making. For some problems requires a high degree of agreement among the participating experts with which arises the need to implement a process of consensus in decision-making in group, thus adding a new phase with the aim of obtaining a high level of agreement among the experts.

It can be said that the consensus is a process of group discussion and iterative is coordinated by a **Moderator** that helps the experts of the problem to bring their views to a general solution accepted by the group. The consensus has been classically defined as the total and unanimous agreement of all the experts involved in the problem, with the passage of time, has been watered down the concept of consensus (*Soft Consensus*) and there have been proposed diffuse measurements that offer a major flexibility to express a vague measurement how the consensus is.

The consensus process consists of four main phases:

### 2.1.1 Expression and collection of preferences

At this stage each expert  $e_i$  expresses a preference on the alternatives group  $X$  by a relationship of fuzzy preference. A formal definition of the assessment of alternatives is as follows:  $X = \{x_1, \dots, x_n\}$ , with ( $n \geq 2$ ) the finite set of alternatives on which a set  $E = \{e_1, \dots, e_m\}$ , with ( $m \geq 2$ ) of experts must provide their preferences. Each expert  $e_i$  shall give its opinion on the set  $X$  via a relationship of fuzzy preference  $P_i : X \times X \rightarrow [0, 1]$ . The special feature of the model is that all the experts  $E$  will not be the same for the different rounds of the process of consensus, in this there may be more or less experts each time.

### 2.1.2 Determination of the degree of consensus

After that each expert in the instant  $t$  has determined their preferences on the set of alternatives is necessary to calculate the degree of consensus  $CR$  which will subsequently be used as a control data to know if it has been finished with rounds of consensus. This process consists of the following steps:

1. **Calculation of matrix of similarity between experts  $SM_{ij}$ :** For each pair of experts  $(e_i, e_j)$ , should calculate how close are their views through the matrix of similarity  $SM_{ij}$ . They will be used for any similarity function such as those presented in [13, 17] for each preference expressed by the experts  $e_i(P_i)$  and  $e_j(P_j)$  for each pair of alternatives  $(x_l, x_k)$  that are present in the decision problem.
2. **Calculation of the matrix of non-dynamic consensus  $CM$ :** Once calculated all the similarity matrices is appropriate to calculate the matrix of consensus for the non-dynamic round  $t$ . This array is aimed at calculating the degree of agreement that exist in the current round among the experts who have participated in the same. This will be used for any operator of aggregation,  $Agg$ , Such as the one presented in [13, 14] and that will result in the array of non-dynamic consensus  $CM$ . After obtaining the matrix  $CM$ , item is the calculation of consensus which will be done in 3 levels [25, 26]:

- (a) Consensus at the level of pairs of alternatives, which is obtained as,

$$cp^{lk} = cm^{lk}, \forall l, k = 1, \dots, n \wedge l \neq k$$

where  $cp^{lk}$  represents the agreement reached on the pair of alternatives  $(x_l, x_k)$ , which is obtained directly from the matrix  $CM$ .

- (b) Consensus at the level of alternatives,

$$ca^l = \phi(cp^{l1}, \dots, cp^{l(l-1)}, cp^{l(l+1)}, \dots, cp^{ln})$$

where  $ca^l$  represents the agreement on the alternative  $x_l$ .

- (c) Consensus at the level of preference.

$$cr = \phi(ca^1, \dots, ca^n)$$

where  $cr$  represents the degree of global consensus reached by experts on the current round.

3. **Calculation of the dynamic matrix of consensus  $DMC$ :** As has been said before, you may find that one or more experts are not present in all the rounds of consensus, by this is that you must use dynamic structures which allow us to recall the opinions that these experts were given in previous rounds. To this effect will be used an aggregation operator  $DE$ , Being suitable for use on a t-norm or t-conorm [ 17], as they meet the associative property. The main advantage of using associative aggregation operators is that it is not necessary to recall all the information on the previous rounds and is therefore only necessary to store the information from the previous round  $t - 1$ .

$$DMC_t = \begin{cases} CM_t & t = 1 \\ DE(DMC_{t-1}, CM_t) & t > 1 \end{cases}$$

When you have the dynamic matrix of consensus for the round  $t$  is possible to obtain the value of consensus  $CR$  through an aggregation operator as shown in [13, 14], in the same way shown in the above matrix, with the calculation of the consensus on three levels [ 25].

### 2.1.3 Control of Consensus

At this stage is checked if it has reached a sufficient degree of agreement to end the process of consensus and move on to the process of selection of alternatives or set of these that correspond to the solution of the decision problem. This is compared to the degree of consensus,  $CR$ , with the threshold value of consensus that has been set prior to the start of the rounds of consensus,  $\gamma$ . This process finishes when  $CR > \gamma$ , otherwise it moves on to the next stage.

### 2.1.4 Generation of recommendations

At this stage the moderator calculates the collective preference of the group, denoted by  $P_c$ , by means of the aggregation of individual preferences of each expert. After having obtained the value of  $P_c$  the moderator proceeds to identify experts  $e_i$  with their respective valuations  $p_i^{lk}$  that are more distant of the consensus and it is recommended to those experts the modification of your assessments, either upwards or derives in order to increase the degree of agreement reached in the next round. Each change recommendation consists of a terna  $(e_i, (x_l, x_k), Direction)$ , which indicates that the expert  $e_i$  must modify the evaluation  $p_i^{lk}$  in the direction given for  $Direction \in \{Increase, Decrease\}$ . Below is the process:

1. **Calculation of the matrix of non-dynamic collective preference  $P_c$ :** This calculation is done so that the moderator is able to determine that experts are more separated from the group in the round  $t$  of the consensus process. For this purpose will be used some operator  $Agg$  as presented in [14], particularly for this array will be used the arithmetic mean operator that will be seen below.
2. **Calculation of the matrix of dynamic collective preference  $P_{cd}$ :** After having realized the calculation of the above matrix must be added the preferences that have in the history of experts who have not participated in the round of current consensus. This will be used for an aggregation operator  $DE$  being suitable for use on a t-norm or t-conorm [17] given the associative properties of these. For the calculation of this dynamic matrix will be used the aggregation operator Average. In this way, the array of collective preference dynamics is calculated as follows:

$$P_{cdt} = \begin{cases} P_{ct} & t = 1 \\ DE(P_{cdt-1}, P_{ct}) & t > 1 \end{cases}$$

Then, at this point, the moderator has all the information necessary to identify those experts and  $e_i$  that have participated in the round  $t$  and whose preferences  $p_i^{lk}$  that is furthest from the consensus group, through the use of measures of similarity, with which you can proceed to generate the recommendations in the same way as in the classic model of consensus.

## 2.2 Consensus measures

In the process of measuring the degree of consensus is necessary to have ways to measure how close are the opinions of the experts among themselves. The consensus measures, therefore, have the goal to be an indicator to evaluate how far they are the opinions of a group of expert in function of a unanimous agreement [20] for then to calculate the degree of consensus through aggregation operators.

usually the process of consensus is a rigid process [21] where the consensus measures only take values  $[0,1]$ , with 0 being a value of disagreement or partial and 1 when there is unanimous agreement. It is here where the measures of *soft consensus* take participation in the process since they provide flexibility for this [13].

In function to achieve its goal, the consensus measures must determine the level of match between each of the

opinions of the experts, and this is achieved through measures of similarity.

## 2.3 Similarity measures

The similarity measures [25] in general are forms of comparison between two or more elements, they measure the similarity between these. Mathematically speaking the similarity is the difference in distance between two values, therefore, its use in the measures of consensus is due to the fact that one of the main features of this process is to measure that as distant are the opinions of various experts, and also measure how distant is the opinion of an expert in relation to the collective.

Then, as discussed earlier, the measures of similarity are strongly related to the distance measures. For purposes of application of the model of GDM will notice the distance between two elements  $(x,y)$  belonging to a set  $X$  as  $d(x,y)$ , And at the same time, it will notice the similarity between two items  $(x,y)$  belonging to a set  $X$  as  $SM(x,y)$ .

Below will explain two existing measures of similarity, where each one of them is based on a comparison between two real values in such a way that  $x, and \in [0, 1]$  with which these will be applied on the same elements  $p_i^{lk}$  Two matrix of preference belonging to different experts.

### 2.3.1 Similarity based on Euclidean Distance

As general definition is called Euclidean distance between two points  $A(x_1, y_1)$  and  $B(x_2, y_2)$  from the plane to the line segment joining those points, it is calculated as:

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

in the case studied is intended to calculate the distance between two assessments delivered by the experts, these assessments correspond to actual values by which we will continue to work with points with a single coordinate, therefore, the distance is as:

$$d(A, B) = \sqrt{(x_2 - x_1)^2}$$

to be  $x_1, x_2 \in [0, 1]$  the values to be calculated for the similarity. The above expression coincides with the difference in absolute value between two numbers therefore can be simplified as:

$$d(A, B) = |x_2 - x_1|$$

from now on it is considered to be the previous expression for the calculation of the distance  $d(x_1, x_2)$ . Then the value of the similarity is obtained by expression:

$$SM(x_1, x_2) = 1 - d(x_1, x_2) = 1 - |x_2 - x_1|$$

The above expression will be used for subsequent calculations that are required in the model of the GDM studied.

### 2.3.2 Quadratic similarity

The quadratic similarity is closely related to the Euclidean distance, where in this case the similarity is given by the square of the distance  $d(x_1, x_2)$ :

$$SM(x_1, x_2) = 1 - d(x_1, x_2)^2$$

This measure to provide higher values than the Euclidean distance fails to achieve a greater convergence to the scope of the consensus, however, suffers from the disadvantage of being a measure already unrealistic that the quadratic be applied an excess in the similarity of the two values, regardless of the distance between them.

## 2.4 Aggregation Operators

Aggregation Operations [25, 26] have merge function as a set of elements of information to obtain a representative element of the same and are a very important element in the decision-making process.

There is a large amount of aggregation operators, which range from operators maximum (OR) operators and the minimum (AND), as also the operators based on average [22]. These operators are generally quite used due to its ease of use.

he aggregation operators are very important because they are used in important processes of decision-making as they are: the calculation of the degree of global consensus (relation of preference) obtained in each moment of time on the basis of consensus obtained for each pair of alternatives, to determine the collective preference of the group and also serve to measure that so far are the opinions of the experts in relation to the collective opinion of them.

There will be denoted the aggregation of a set of values  $X = \{x_1, \dots, x_n\}$  as  $\phi(x_1, \dots, x_n)$ , where the main functions are the following:

- Continuity:  $\phi$  it is a continuous function for each of its variables.
- Commutativity:  $\phi(x, y) = \phi(y, x), \forall x, y \in X$
- Associativity:  $\phi(x, \phi(y, z)) = \phi(\phi(x, y), z), \forall x, y, z \in X$
- Monotony: Si  $x \leq z$ , then  $\phi(x, y) = \phi(z, y)$
- Idempotence:  $\phi(x, \dots, x) = x, \forall x \in X$

Below are different aggregation operators:

### 2.4.1 Arithmetic mean

Popularly known as average, the arithmetic mean  $\bar{x}$  of a set of values  $X = \{x_1, \dots, x_n\}$  It is the sum of the values divided by the number of them:

$$\phi(x_1, \dots, x_n) = \bar{x} = \sum_i^n \frac{x_i}{n}$$

the main disadvantage of this operator is that it is easily affected by minimum and maximum values thus making it a unreliable operator as in some cases it may be a measure unrepresentative of the set of values. Some of the features for this operator are:

- Null Offset:  $\sum_i (x_i - \bar{x}) = 0$
- Given a value  $a$ ,  $\sum_i (x_i - a)^2$  is minimal when  $a = \bar{x}$
- For each  $x_i$ ,  $\sum_i^n (x_i + a)/n = \bar{x} + a$
- For each  $x_i$ ,  $\sum_i^n (x_i * a)/n = \bar{x} * a$

In addition the arithmetic mean complies with the properties of continuity and Idempotence, commutativity, monotony, but does not meet the properties of associativity.

### 2.4.2 Geometric Mean

The geometric mean is defined as the n-th root of the product of all the numbers within a set of values  $X$ :

$$\phi(x_1, \dots, x_n) = \bar{x} = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \dots x_n}$$

One of the main advantages of this aggregator, is that it is less sensitive than the arithmetic mean to extreme values within a dataset. However, presents a major disadvantage when some value  $x_i$  is equal to 0, because the result will be equal to 0 independent of the other values of the whole. This operator will be used for the calculations of matrix of consensus which do not require a associative operator based in t-norm and t-conorm.

### 2.4.3 OWA Operators

Given a set of elements  $X = \{x_1, \dots, x_n\}$  and a set of weights  $W = \{w_1, \dots, w_n\}$ , such that  $w_i \in [0, 1]$  and the sum of the elements  $w_i$  It is equal to 1, the associated OWA operator, who will be denoted like  $F$ , is defined as:

$$F(x_1, \dots, x_n) = w_1 b_1 + w_2 b_2 + \dots + w_n b_n$$

where  $b_i$  It is the largest  $i$ -th element of  $X$ , and the main feature of these aggregation operators is that the values to add are sorted from highest to lowest. The OWA operators require a set of weights  $w_i$  for each element, each of which is associated with the  $i$ -th largest element  $b_i$  (not necessarily  $x_i$ ). The OWA operators allow you to define a measure of optimism (orness) depending on the set of weights used,  $W$ :

$$orness(W) = \frac{1}{n-1} \sum_{i=1}^n ((n-i)w_i)$$

This measure is within the interval  $[0,1]$ , and determines the degree to which the aggregation is next to the OR operator (maximum), being 1 when using the own OR operator. If the value  $orness(W) = 0$ , this means that we are using the AND operator. On the other hand, when you use the arithmetic mean is obtained a value of  $orness(W) = 0.5$  (Degree of optimism neutral, mid-way between AND and OR).

Similarly defined a measure of pessimism of the OWA operator used, such as:

$$andness(W) = 1 - orness(W) = 1 - \frac{1}{n-1} \sum_{i=1}^n ((n-i)w_i)$$

### 2.4.4 OR operator (Maximum)

It is simply to consider such as the aggregation of a set of values the maximum of the set. This operator represents a particular case of OWA operator, where we consider  $w_1 = 1$  and  $w_i = 0$  for  $i \neq 1$ :

$$\phi(x_1, \dots, x_n) = F^*(x_1, \dots, x_n) = \max_i[x_i]$$

The notion of this operator as a particular case of the family of OWA allows operators to make an extension of the same, giving rise to the so-called operator  $k$ -maximum, where taking  $w_k = 1$ , ( $1 \leq k \leq n$ ) and  $w_i = 0$ , ( $i \neq k$ ) it is obtained as the value of aggregation the  $k$ -th largest value of the set.

Despite being a very simple way to add a series of values, the OR operator is unrepresentative when there is a large variability or dispersion between those values. The OR operator corresponds to a feature  $t$ -conorm therefore complies with the commutative property, associativity, among others.

### 2.4.5 AND operator (Minimum)

In a similar way to the OR operator, in this case we consider the minimum value as aggregation of the whole. The AND operator is equivalent to the particular case of OWA, where  $w_i = 0$  for  $i \neq n$  and  $w_n = 1$ .

$$\phi(x_1, \dots, x_n) = F^+(x_1, \dots, x_n) = \min_i[x_i]$$

Like the OR operator, this form of aggregation may be unrepresentative when we have very different values between if. Also, the notion of this operator as a particular case of the family of operators OWA lets us extend it, giving rise to the so-called operator  $k$ -minimum, where taking  $w_{n-k+1} = 1$ , ( $1 \leq k \leq n$ ) and  $w_i = 0$ , ( $i \neq n - k + 1$ ) It is obtained as the value of aggregation the  $k$ -th smallest value of the set. The AND operator corresponds to a feature  $t$ -norm, therefore complies with the commutative property, associativity, among others.

### 2.4.6 Laplace Operator

This operator, also known as *Normative approach*, It is a operator of classic decision widely used both in decision-making as in statistical processes. Its use is simple, is to calculate the value representative of each vector  $c_i$  in the following way:

$$G_N(c_i) = \frac{1}{q} \sum_{j=1}^q c_{ij}$$

following the two previous operators, the final value added is obtained as:

$$\phi(c_1, \dots, c_n) = G_N(C) = \max_i G_N(c_i)$$

In summary, it is possible to appreciate that the Laplace operator is in fact a combination of the arithmetical mean with the operator OR. The degree of optimism (*orness*) for the Laplace operator is 0.5.

### 2.4.7 Average Operator

The average operator, for  $p \in [0, 1]$  it is expressed as [23]:

$$Med_p(x, y) = \begin{cases} \max(x, y) & x, y \leq p \\ \min(x, y) & x, y \geq p \\ p & \text{in other case} \end{cases}$$

where agreement with Dubois and Prade [ 24], this is the only media associative operator between the maximum and the minimum to be a operator of aggregation based on  $t$ -norm and  $t$ -conorm. This operator will be used for the calculations of dynamic matrix in where it is appropriate to use an associative operator based in  $t$ -norm and  $t$ -conorm [17].

### 3 The Higher Council

The Higher Council of the Technological Metropolitan University is the collegiate organism of major hierarchy and it is entrusted to carry out the decision making of major proportions inside the university. The Higher Council is integrated by nine members with right to vote, also, there take part with right to voice two representatives of the students and one of not academic officials. The members with right to vote are the following ones:

1. Three councillors appointed by the President of the Republic of Chile.
2. Five advisers chosen by the academic body, in accordance with the regulation of election that the proper Higher Council fixes.
3. The Rector of the University, who presides at the Higher Council.

### 4 Case of Application

The group decision making is devoted to select an option where most feel comfortable with it, it is here where the consensus process works. As already mentioned above, it seeks to reach compliance more than it would be with the classical group decision-making process, since it make use of preferences in each option seeks to align the alternatives to a mostly accepted common, even more so when you access the opinions of any missing members through the use of a dynamic decision structure.

To test the functioning of the decision-making model studied will be the following:

- There is a problem  $A$  which will consist of four possible solutions corresponding to the set:  $X = \{x_1, x_2, x_3, x_4, x_5\}$ .
- The President of the Council will act as the moderator of the session's decision.
- The eight remaining members with vote of the Council will be typified as:  $Expert1, Expert2, \dots, Expert8$ .
- The three remaining members without vote of the Council will not be considered for purposes of the example.
- Each expert will express their views through a matrix of fuzzy preference

$$P_i = \begin{pmatrix} - & \dots & p_i^{15} \\ \vdots & \ddots & \vdots \\ p_i^{51} & \dots & - \end{pmatrix}$$

where each valuation  $p_i^{lk}$  it will express the degree of preference of the alternative  $x_l$  on  $x_k$  if  $p_i^{lk} > 0.5$  o la preferencia de  $x_k$  on  $x_l$  si  $p_i^{lk} < 0.5$ , any valuation  $p_i^{lk} = 0.5$  it indicates indifference between the pair of alternatives.

- The degree of preference expressed by each expert can never be equal to 0.
- Each recommendation generated by the moderator will be accepted by the experts, where if it is recommended to increase a valuation  $p_i^{lk}$  this must be done with a value of 0.1 , on the other hand, if it is recommended to decrease a valuation  $p_i^{lk}$  this must be done with a value of 0.1 .
- Expert valuations have the same degree of importance among themselves.
- It is considered that an agreement was reached when the degree of consensus is greater than or equal to threshold  $\gamma = 0.85$ .
- The maximum number of rounds permitted in the consensus process will be 10.
- All the recommendations expressed by the moderator will be of *Increase*.

For all practical purposes, in each round will be displayed only arrays of preferences of each expert and the calculation of the consensus. For the rounds where  $t > 1$  The recommendations made by the moderator on the experts will be reflected in blue numbers within their matrix of preferences, in addition, for the calculation of matrix of dynamic collective preference is to be made use of average operator with  $p = 0.7$ .

#### 4.1 First Round of Decision

The following are the preference of arrays generated by the eight participating experts of the problem for  $t = 1$ , where the positions for  $i = j$  they will not value since it represents the preference over the same alternative.

$$P_1(i, j) = \begin{pmatrix} - & 0.681 & 0.754 & 0.855 & 0.423 \\ 0.319 & - & 0.454 & 0.599 & 0.684 \\ 0.246 & 0.546 & - & 0.445 & 0.781 \\ 0.145 & 0.401 & 0.555 & - & 0.325 \\ 0.577 & 0.316 & 0.219 & 0.675 & - \end{pmatrix}$$

$$P_2(i, j) = \begin{pmatrix} - & 0.458 & 0.328 & 0.489 & 0.589 \\ 0.542 & - & 0.454 & 0.875 & 0.498 \\ 0.672 & 0.546 & - & 0.654 & 0.369 \\ 0.511 & 0.125 & 0.346 & - & 0.588 \\ 0.411 & 0.502 & 0.631 & 0.412 & - \end{pmatrix}$$

$$P_3(i, j) = \begin{pmatrix} - & 0.787 & 0.687 & 0.985 & 0.232 \\ 0.213 & - & 0.487 & 0.699 & 0.777 \\ 0.313 & 0.513 & - & 0.359 & 0.784 \\ 0.015 & 0.301 & 0.641 & - & 0.369 \\ 0.768 & 0.223 & 0.216 & 0.631 & - \end{pmatrix}$$

$$P_4(i, j) = \begin{pmatrix} - & 0.565 & 0.874 & 0.988 & 0.365 \\ 0.435 & - & 0.588 & 0.988 & 0.125 \\ 0.126 & 0.412 & - & 0.369 & 0.458 \\ 0.012 & 0.012 & 0.631 & - & 0.125 \\ 0.635 & 0.875 & 0.542 & 0.875 & - \end{pmatrix}$$

$$P_5(i, j) = \begin{pmatrix} - & 0.588 & 0.547 & 0.258 & 0.688 \\ 0.412 & - & 0.588 & 0.588 & 0.784 \\ 0.453 & 0.412 & - & 0.329 & 0.799 \\ 0.742 & 0.412 & 0.671 & - & 0.455 \\ 0.312 & 0.216 & 0.201 & 0.545 & - \end{pmatrix}$$

$$P_6(i, j) = \begin{pmatrix} - & 0.658 & 0.655 & 0.729 & 0.652 \\ 0.342 & - & 0.499 & 0.678 & 0.429 \\ 0.345 & 0.501 & - & 0.421 & 0.741 \\ 0.271 & 0.322 & 0.579 & - & 0.224 \\ 0.348 & 0.571 & 0.259 & 0.776 & - \end{pmatrix}$$

$$P_7(i, j) = \begin{pmatrix} - & 0.566 & 0.788 & 0.741 & 0.333 \\ 0.434 & - & 0.566 & 0.688 & 0.684 \\ 0.212 & 0.434 & - & 0.333 & 0.688 \\ 0.259 & 0.312 & 0.667 & - & 0.255 \\ 0.667 & 0.316 & 0.312 & 0.745 & - \end{pmatrix}$$

$$P_8(i, j) = \begin{pmatrix} - & 0.588 & 0.652 & 0.985 & 0.599 \\ 0.412 & - & 0.308 & 0.366 & 0.874 \\ 0.348 & 0.692 & - & 0.332 & 0.333 \\ 0.015 & 0.634 & 0.669 & - & 0.455 \\ 0.401 & 0.126 & 0.667 & 0.545 & - \end{pmatrix}$$

#### 4.1.1 Calculation of the matrix of non-dynamic Consensus $CM$

Once calculated all the matrix of similarity between all the participating experts of the round  $t = 1$  It is appropriate to calculate the matrix of consensus not dynamics

which, as explained before, has aimed at obtaining the degree of agreement among the experts of the current round. For this purpose will be using the aggregation operator Geometric Mean. The matrix of non-dynamic consensus  $CM$  it appears next:

$$CM_1 = \begin{pmatrix} - & 0.883 & 0.795 & 0.654 & 0.788 \\ & - & 0.890 & 0.765 & 0.683 \\ & & - & 0.882 & 0.755 \\ & & & - & 0.812 \\ & & & & - \end{pmatrix}$$

Then proceed with the calculation of the degree of global consensus for the round  $t = 1$ :

1. Consensus at the level of pair of alternatives:

$$CM_1 = \begin{pmatrix} - & 0.883 & 0.795 & 0.654 & 0.788 \\ & - & 0.890 & 0.765 & 0.683 \\ & & - & 0.882 & 0.755 \\ & & & - & 0.812 \\ & & & & - \end{pmatrix}$$

2. Consensus at the level of alternatives:

$$\begin{aligned} ca^1 &= \phi(0.881, 0.795, 0.654, 0.788) = 0.776 \\ ca^2 &= \phi(0.883, 0.890, 0.765, 0.683) = 0.801 \\ ca^3 &= \phi(0.795, 0.890, 0.882, 0.755) = 0.828 \\ ca^4 &= \phi(0.654, 0.765, 0.882, 0.812) = 0.774 \\ ca^5 &= \phi(0.788, 0.683, 0.755, 0.812) = 0.758 \end{aligned}$$

3. Consensus at the level of relationship of preference:

$$cr = \phi(0.776, 0.801, 0.828, 0.774, 0.758) = 0.787$$

As no consensus was reached is passed to the second round of decision.

## 4.2 Second Round of Decision

For this round will not be counted with the participation of the Expert3 and it will therefore not be considered in further calculations, however, will be expressed his preference in red color matrix.

$$P_1(i, j) = \begin{pmatrix} - & 0.681 & 0.754 & 0.855 & \mathbf{0.523} \\ 0.319 & - & \mathbf{0.554} & \mathbf{0.699} & 0.684 \\ 0.246 & 0.446 & - & 0.445 & 0.781 \\ 0.145 & 0.301 & 0.555 & - & \mathbf{0.425} \\ 0.477 & 0.316 & 0.219 & 0.575 & - \end{pmatrix}$$

$$P_2(i, j) = \begin{pmatrix} - & \mathbf{0.558} & \mathbf{0.428} & \mathbf{0.589} & 0.589 \\ 0.442 & - & \mathbf{0.554} & 0.875 & \mathbf{0.598} \\ 0.572 & 0.446 & - & 0.654 & \mathbf{0.469} \\ 0.411 & 0.125 & 0.346 & - & 0.588 \\ 0.411 & 0.402 & 0.531 & 0.412 & - \end{pmatrix}$$

$$P_3(i, j) = \begin{pmatrix} - & 0.787 & 0.687 & 0.985 & 0.332 \\ 0.213 & - & 0.587 & 0.699 & 0.777 \\ 0.313 & 0.413 & - & 0.459 & 0.784 \\ 0.015 & 0.301 & 0.541 & - & 0.369 \\ 0.668 & 0.223 & 0.216 & 0.631 & - \end{pmatrix}$$

$$P_4(i, j) = \begin{pmatrix} - & 0.665 & 0.874 & 0.988 & 0.465 \\ 0.335 & - & 0.588 & 0.988 & 0.225 \\ 0.126 & 0.412 & - & 0.469 & 0.558 \\ 0.012 & 0.012 & 0.531 & - & 0.225 \\ 0.535 & 0.775 & 0.442 & 0.775 & - \end{pmatrix}$$

$$P_5(i, j) = \begin{pmatrix} - & 0.688 & 0.647 & 0.358 & 0.688 \\ 0.312 & - & 0.588 & 0.688 & 0.784 \\ 0.353 & 0.412 & - & 0.429 & 0.799 \\ 0.642 & 0.312 & 0.571 & - & 0.455 \\ 0.312 & 0.216 & 0.201 & 0.545 & - \end{pmatrix}$$

$$P_6(i, j) = \begin{pmatrix} - & 0.658 & 0.755 & 0.829 & 0.652 \\ 0.342 & - & 0.499 & 0.778 & 0.529 \\ 0.245 & 0.501 & - & 0.421 & 0.741 \\ 0.171 & 0.222 & 0.579 & - & 0.324 \\ 0.348 & 0.471 & 0.259 & 0.676 & - \end{pmatrix}$$

$$P_7(i, j) = \begin{pmatrix} - & 0.667 & 0.788 & 0.841 & 0.433 \\ 0.333 & - & 0.566 & 0.688 & 0.684 \\ 0.212 & 0.434 & - & 0.433 & 0.688 \\ 0.159 & 0.312 & 0.567 & - & 0.355 \\ 0.567 & 0.316 & 0.312 & 0.645 & - \end{pmatrix}$$

$$P_8(i, j) = \begin{pmatrix} - & 0.688 & 0.752 & 0.985 & 0.599 \\ 0.312 & - & 0.408 & 0.466 & 0.874 \\ 0.248 & 0.592 & - & 0.432 & 0.433 \\ 0.015 & 0.534 & 0.568 & - & 0.455 \\ 0.401 & 0.126 & 0.567 & 0.545 & - \end{pmatrix}$$

#### 4.2.1 Calculation of the dynamic matrix of Consensus $DMC$

In this second round, and in the post, it must calculate the dynamic array  $DMC$  through the use of an aggregation operator being suitable for use on a t-norm or t-conorm. The average operator is to be used with  $p = 0.9$ , therefore, the dynamic matrix of consensus for the round  $t = 2$  it is the next:

$$DMC_2 = \begin{pmatrix} - & 0.883 & 0.834 & 0.713 & 0.882 \\ & - & 0.890 & 0.795 & 0.734 \\ & & - & 0.882 & 0.811 \\ & & & - & 0.857 \\ & & & & - \end{pmatrix}$$

Then proceed with the calculation of the degree of global consensus for the round  $t = 2$ :

1. Consensus at the level of pair of alternatives:

$$DMC_2 = \begin{pmatrix} - & 0.883 & 0.834 & 0.713 & 0.882 \\ & - & 0.890 & 0.795 & 0.734 \\ & & - & 0.882 & 0.811 \\ & & & - & 0.857 \\ & & & & - \end{pmatrix}$$

2. Consensus at the level of alternatives:

$$\begin{aligned} ca^1 &= \phi(0.883, 0.834, 0.713, 0.882) = 0.825 \\ ca^2 &= \phi(0.883, 0.890, 0.795, 0.734) = 0.823 \\ ca^3 &= \phi(0.834, 0.890, 0.882, 0.811) = 0.854 \\ ca^4 &= \phi(0.713, 0.795, 0.882, 0.857) = 0.809 \\ ca^5 &= \phi(0.882, 0.734, 0.811, 0.857) = 0.819 \end{aligned}$$

3. Consensus at the level of relationship of preference:

$$cr = \phi(0.825, 0.823, 0.854, 0.809, 0.819) = 0.826$$

As no consensus was reached is passed to the Third Round of decision.

### 4.3 Third Round of Decision

For this round will not be counted with the participation of the Expert1 and the Expert3 for which shall not be considered in further calculations, however, will be expressed its preference in red with the recommendations expressed by the moderator.

$$P_1(i, j) = \begin{pmatrix} - & 0.681 & 0.754 & 0.855 & 0.623 \\ 0.319 & - & 0.554 & 0.699 & 0.684 \\ 0.246 & 0.446 & - & 0.545 & 0.781 \\ 0.145 & 0.301 & 0.455 & - & 0.425 \\ 0.377 & 0.316 & 0.219 & 0.575 & - \end{pmatrix}$$

$$P_2(i, j) = \begin{pmatrix} - & 0.658 & 0.528 & 0.689 & 0.589 \\ 0.342 & - & 0.554 & 0.875 & 0.698 \\ 0.472 & 0.446 & - & 0.654 & 0.569 \\ 0.311 & 0.125 & 0.346 & - & 0.588 \\ 0.411 & 0.302 & 0.431 & 0.412 & - \end{pmatrix}$$

$$P_3(i, j) = \begin{pmatrix} - & 0.787 & 0.687 & 0.985 & 0.332 \\ 0.213 & - & 0.587 & 0.699 & 0.777 \\ 0.313 & 0.413 & - & 0.459 & 0.784 \\ 0.015 & 0.301 & 0.541 & - & 0.369 \\ 0.668 & 0.223 & 0.216 & 0.631 & - \end{pmatrix}$$

$$P_4(i, j) = \begin{pmatrix} - & 0.665 & 0.874 & 0.988 & \mathbf{0.565} \\ 0.335 & - & 0.588 & 0.988 & \mathbf{0.325} \\ 0.126 & 0.412 & - & 0.469 & \mathbf{0.658} \\ 0.012 & 0.012 & 0.531 & - & \mathbf{0.325} \\ 0.435 & 0.675 & 0.342 & 0.675 & - \end{pmatrix}$$

$$P_5(i, j) = \begin{pmatrix} - & 0.688 & \mathbf{0.747} & \mathbf{0.458} & 0.688 \\ 0.312 & - & 0.588 & 0.688 & 0.784 \\ 0.253 & 0.412 & - & \mathbf{0.529} & 0.799 \\ 0.542 & 0.312 & 0.471 & - & 0.455 \\ 0.312 & 0.216 & 0.201 & 0.545 & - \end{pmatrix}$$

$$P_6(i, j) = \begin{pmatrix} - & 0.658 & 0.755 & 0.829 & 0.652 \\ 0.342 & - & \mathbf{0.599} & 0.778 & \mathbf{0.629} \\ 0.245 & 0.401 & - & \mathbf{0.521} & 0.741 \\ 0.171 & 0.222 & 0.479 & - & \mathbf{0.424} \\ 0.348 & 0.371 & 0.259 & 0.576 & - \end{pmatrix}$$

$$P_7(i, j) = \begin{pmatrix} - & 0.667 & 0.788 & 0.841 & \mathbf{0.533} \\ 0.333 & - & 0.566 & 0.688 & 0.684 \\ 0.212 & 0.434 & - & \mathbf{0.533} & 0.688 \\ 0.159 & 0.312 & 0.467 & - & \mathbf{0.455} \\ 0.467 & 0.316 & 0.312 & 0.545 & - \end{pmatrix}$$

$$P_8(i, j) = \begin{pmatrix} - & 0.688 & 0.752 & 0.985 & 0.599 \\ 0.312 & - & \mathbf{0.508} & \mathbf{0.566} & 0.874 \\ 0.248 & 0.492 & - & \mathbf{0.532} & \mathbf{0.533} \\ 0.015 & 0.434 & 0.468 & - & 0.455 \\ 0.401 & 0.126 & 0.467 & 0.545 & - \end{pmatrix}$$

### 4.3.1 Calculation of the dynamic array of Consensus $DMC$

In this third round, it must calculate the dynamic array  $DMC$  through the use of an aggregation operator being suitable for use on a t-norm or t-conorm. The average operator is to be used with  $p = 0.9$ , therefore, the dynamic matrix of consensus for the round  $t = 3$  it is the next:

$$DMC_3 = \begin{pmatrix} - & 0.883 & 0.870 & 0.745 & 0.882 \\ & - & 0.890 & 0.808 & 0.767 \\ & & - & 0.882 & 0.872 \\ & & & - & 0.857 \\ & & & & - \end{pmatrix}$$

Then proceed with the calculation of the degree of global consensus for the round  $t = 3$ :

1. Consensus at the level of pair of alternatives:

$$DMC_3 = \begin{pmatrix} - & 0.883 & 0.870 & 0.745 & 0.882 \\ & - & 0.890 & 0.808 & 0.767 \\ & & - & 0.882 & 0.872 \\ & & & - & 0.857 \\ & & & & - \end{pmatrix}$$

2. Consensus at the level of alternatives:

$$\begin{aligned} ca^1 &= \phi(0.883, 0.870, 0.745, 0.882) = 0.843 \\ ca^2 &= \phi(0.883, 0.890, 0.808, 0.767) = 0.835 \\ ca^3 &= \phi(0.870, 0.890, 0.882, 0.872) = 0.878 \\ ca^4 &= \phi(0.745, 0.808, 0.882, 0.857) = 0.821 \\ ca^5 &= \phi(0.882, 0.767, 0.874, 0.857) = 0.844 \end{aligned}$$

3. Consensus at the level of relationship of preference:

$$cr = \phi(0.843, 0.835, 0.878, 0.821, 0.844) = 0.844$$

As no consensus was reached is passed to the fourth round of decision.

## 4.4 Fourth Round of Decision

For this round will not be counted with the participation of the Expert1 and the Expert3 for which shall not be considered in further calculations, however, will be expressed its preference in red with the recommendations expressed by the moderator.

$$P_1(i, j) = \begin{pmatrix} - & \mathbf{0.681} & \mathbf{0.754} & \mathbf{0.855} & \mathbf{0.623} \\ \mathbf{0.319} & - & \mathbf{0.554} & \mathbf{0.699} & \mathbf{0.684} \\ \mathbf{0.246} & \mathbf{0.446} & - & \mathbf{0.545} & \mathbf{0.781} \\ \mathbf{0.145} & \mathbf{0.301} & \mathbf{0.455} & - & \mathbf{0.425} \\ \mathbf{0.377} & \mathbf{0.316} & \mathbf{0.219} & \mathbf{0.575} & - \end{pmatrix}$$

$$P_2(i, j) = \begin{pmatrix} - & \mathbf{0.758} & \mathbf{0.628} & \mathbf{0.789} & \mathbf{0.689} \\ 0.242 & - & \mathbf{0.654} & 0.875 & 0.698 \\ 0.372 & 0.346 & - & 0.654 & \mathbf{0.669} \\ 0.211 & 0.125 & 0.346 & - & 0.588 \\ 0.311 & 0.302 & 0.331 & 0.412 & - \end{pmatrix}$$

$$P_3(i, j) = \begin{pmatrix} - & \mathbf{0.787} & \mathbf{0.687} & \mathbf{0.985} & \mathbf{0.332} \\ \mathbf{0.213} & - & \mathbf{0.587} & \mathbf{0.699} & \mathbf{0.777} \\ \mathbf{0.313} & \mathbf{0.413} & - & \mathbf{0.459} & \mathbf{0.784} \\ \mathbf{0.015} & \mathbf{0.301} & \mathbf{0.541} & - & \mathbf{0.369} \\ \mathbf{0.668} & \mathbf{0.223} & \mathbf{0.216} & \mathbf{0.631} & - \end{pmatrix}$$

$$P_4(i, j) = \begin{pmatrix} - & \mathbf{0.765} & 0.874 & 0.988 & \mathbf{0.665} \\ 0.235 & - & 0.588 & 0.988 & \mathbf{0.425} \\ 0.126 & 0.412 & - & \mathbf{0.569} & \mathbf{0.758} \\ 0.012 & 0.012 & 0.431 & - & \mathbf{0.425} \\ 0.335 & 0.575 & 0.242 & 0.575 & - \end{pmatrix}$$

$$P_5(i, j) = \begin{pmatrix} - & 0.688 & 0.747 & \mathbf{0.558} & 0.688 \\ 0.312 & - & 0.588 & 0.688 & 0.784 \\ 0.253 & 0.412 & - & \mathbf{0.629} & 0.799 \\ 0.442 & 0.312 & 0.371 & - & 0.455 \\ 0.312 & 0.216 & 0.201 & 0.545 & - \end{pmatrix}$$

$$P_6(i, j) = \begin{pmatrix} - & \mathbf{0.758} & 0.755 & 0.829 & 0.652 \\ 0.242 & - & 0.599 & 0.778 & \mathbf{0.729} \\ 0.245 & 0.401 & - & \mathbf{0.621} & 0.741 \\ 0.171 & 0.222 & 0.379 & - & \mathbf{0.524} \\ 0.348 & 0.271 & 0.259 & 0.476 & - \end{pmatrix}$$

$$P_7(i, j) = \begin{pmatrix} - & \mathbf{0.767} & 0.788 & 0.841 & \mathbf{0.633} \\ 0.233 & - & \mathbf{0.667} & 0.688 & 0.684 \\ 0.212 & 0.333 & - & \mathbf{0.633} & 0.688 \\ 0.159 & 0.312 & 0.367 & - & 0.455 \\ 0.367 & 0.316 & 0.312 & 0.545 & - \end{pmatrix}$$

$$P_8(i, j) = \begin{pmatrix} - & 0.688 & 0.752 & 0.985 & \mathbf{0.699} \\ 0.312 & - & \mathbf{0.608} & \mathbf{0.667} & 0.874 \\ 0.248 & 0.392 & - & \mathbf{0.632} & \mathbf{0.633} \\ 0.015 & 0.333 & 0.368 & - & 0.455 \\ 0.301 & 0.126 & 0.367 & 0.545 & - \end{pmatrix}$$

#### 4.4.1 Calculation of the dynamic matrix of Consensus *DMC*

In this fourth round, it must calculate the dynamic array *DMC* through the use of an aggregation operator being suitable for use on a t-norm or t-conorm. The average operator is to be used with  $p = 0.9$ , therefore, the dynamic matrix of consensus for the round  $t = 4$  it is the next:

$$DMC_4 = \begin{pmatrix} - & 0.883 & 0.870 & 0.806 & 0.882 \\ & - & 0.890 & 0.843 & 0.818 \\ & & - & 0.882 & 0.874 \\ & & & - & 0.857 \\ & & & & - \end{pmatrix}$$

Then proceed with the calculation of the degree of global consensus for the round  $t = 4$ :

1. Consensus at the level of pair of alternatives:

$$DMC_4 = \begin{pmatrix} - & 0.883 & 0.870 & 0.806 & 0.882 \\ & - & 0.890 & 0.843 & 0.818 \\ & & - & 0.882 & 0.874 \\ & & & - & 0.857 \\ & & & & - \end{pmatrix}$$

2. Consensus at the level of alternatives:

$$\begin{aligned} ca^1 &= \phi(0.883, 0.870, 0.806, 0.882) = 0.860 \\ ca^2 &= \phi(0.883, 0.890, 0.843, 0.818) = 0.858 \\ ca^3 &= \phi(0.870, 0.890, 0.882, 0.874) = 0.879 \\ ca^4 &= \phi(0.806, 0.843, 0.882, 0.857) = 0.847 \\ ca^5 &= \phi(0.882, 0.818, 0.874, 0.857) = 0.857 \end{aligned}$$

3. Consensus at the level of relationship of preference:

$$cr = \phi(0.860, 0.858, 0.879, 0.847, 0.857) = 0.860$$

Then, as it can see, consensus is achieved

## 5 Conclusions

### 5.1 Generals

The decision-making, as was said at the beginning, it is a fundamental and basic process that every human being performs throughout his life, is the process by which a choice between the options or ways to deal with various situations in life in different contexts, at work, family, sentimental, business, using methodologies that provides the administration. The decision-making is basically in the election of an option among the available in order to solve a current problem or potential.

The decision-making at group level is a complex process that requires that the vast majority of the parties involved in the decision to reach an agreement on the same, which can become a long and cumbersome process.

The use of models that support the decision-making gives a very useful tool which allows you to shorten your times of decision and at the same time, they help make a better decision in the face of a problem. The model studied in this work of titling has the great advantage of being able to "remember" the views given in each moment of time, thus making it a stable model.

Also, as could be seen, there are numerous aggregation operators which have different effects, for example, some of them can make the process of consensus is short at the expense of a less than optimal solution, in contrast, other operators slower to reach consensus may obtain optimal solutions more than their peers faster.

There are various models of GDM, some focused on the qualitative assessments and other in the quantitative, with homogeneous and heterogeneous information, based in different contexts, but all of them point toward the same goal which is, take a right decision when faced with a problem or situation.

Finally, it emphasizes that the use of the model under consideration would greatly assist in the decision-making processes that are currently being implemented in the

Superior Council of the Technology Metropolitan University of the State of Chile, making these processes are fast and reliable; that lead them to make the best decisions. It is intended to be a support system to the University, not only within this area, but also in the different areas that make up the organization, as it is normal that within each group of people at some point it will need to make a group decision, where again the technologies are present to attend.

## 5.2 Analysis of Results

By way of analysis of the results obtained in the process, it was decided to use the operator of Geometric Mean due to the fact that they have a lower sensitivity to extreme values within a set of values, as well as taking the restriction that the assessments delivered by the experts could not be equal to 0 eliminated the drawback to this operator of aggregation. At the same time, the average operator was used for all calculations requiring dynamism to "remember" the previous opinions expressed by the experts.

The results seen show that it is possible to achieve the consensus of a rapid and satisfactory manner within a group of people, where the degree of agreement between all of them is a minimum stay of one 85%, which is better than the current process of "most wins," that is, an agreement of most of the 50%.

The decision-making process of the Higher Council that is currently being carried out is slow and sometimes unsatisfactory, it can take a long time close topic important due to the fact that it must meet all the members to discuss the issues and reach an agreement. What it proposes this model is to make this process something reliable, successful, efficient, fast, in order to arrive at the best alternative within the proposals in the problem. With the use of the model of studied TDG, major "freedom" is granted since it is possible that in certain decision round it does not appear any of the members of the Advice with which, in the current process, it would provoke a problem since the opinion would be missing with regard to the topic that it is talking each other, not this way with the use of the model, since, on having been based on dynamic structures, it gives the possibility of maintaining the opinions in the time, in order to which if in some given moment some of the members was going so far as to be missing of the Advice, its opinion is always present.

In the first round of decision, was obtained a consensus value of the 78.7% which was not satisfying the agreement condition, for which it passed to the stage of recommendations where the moderator, in this case the Rec-

tor proceeds, based on the collective opinion, to identify those experts who were removed from the group opinion, recommending to them to increase or to decrease certain opinions. Already for the second round the agreement grade increased one 82.6%, where once again was not met the threshold of consensus building, for which was repeated the process of recommendations coming as well to a third round of decision where the degree of agreement under a value of 84.4% and finally in the fourth round of the degree of consensus agreement passed the threshold with a value of 86.0%. Thereafter it is possible to determine the alternative that most pleases the group of people through some aggregation operator such as those seen in [13, 14] on the consensus at the level of alternatives. In the case studied, using the OR operator on the consensus at the level of alternatives in the fourth round of decision, is that the alternative  $x_3$  it would be the selected to resolve the problem. Once has been selected the alternative that would solve the problem, it is appropriate to run it and follow up with the same in order to test its operation and verify that they comply with its purpose, if this is not fulfilled as expected, would be to generate new alternatives or modifications of the above which would give way to a new decision-making process in group with consensus process based on dynamic decision-making structures.

## 5.3 Future Work

As future work for GDM developed model is the practical application of the model in real situations, not only within the studied field, but they also extend it to areas where it is required a consensual group decision. A proposal, currently at the moment in which a group of people within a company should talk about something and make a decision about this, typically, gather in a meeting room, discuss the issue and reach some sort of conclusion on the basis of the opinions proposed the participants of these meetings, will propose you then create a computer system that applies the model of GDM studied in such a way that these meeting rooms will become "smart meeting rooms". A smart meeting room is a support that allows solutions to reach optimal a way to quickly and accurately, where every participant in this meeting against a computational device which displays options for solution to a given problem that the expert may enter such alternative estimations and then to perform the process of consensus in a computerized manner. Another interesting point is the change the model so that it supports different priorities among the participating experts, where certain views will have more "weight" within the process of consensus. Another option is that the model supports

more alternatives in any instant of the consensus process, or that even at the time that you reach consensus, able to choose more than one alternative which complement each other.

Trends in group decision making are aimed at broadening the spectrum of implementation of the processes of these in order to bring them to different areas of application, not only in the sphere of business. What is sought is to reach the best solutions in an efficient manner, optimizing resources and time to get a greater benefit.

## 6 References

- [1] R. Duncan and H. Raiffa. Games and Decision. *Introduction and Critical Survey*. Dover Publications, 1985.
- [2] S. Rios, C. Bielza, and A. Mateos. Fundamentos de los Sistemas de Ayuda a la Decisión. Ra-Ma, 2002.
- [3] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8:338-353, 1965.
- [4] R. Caballero and G. M. Fernández. Toma de Decisiones con Criterios Múltiples. *Revista de Comunicaciones y Trabajos de ASEPUMA*, 2002.
- [5] C. Romero. Teoría de la Decisión Multicriterio: Conceptos, Técnicas, Aplicaciones. Alianza Universidad, 1993.
- [6] J. Doyle. Prospects for preferences. *Computational Intelligence*, 20(2):111-136, 2004.
- [7] E. Herrera-Viedma, F. Herrera, and F. Chiclana. A consensus model for multiperson decision making with different preference structures. *IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans*, 32(3):394-402, 2002.
- [8] T. Tanino. On Group Decision Making Under Fuzzy Preferences, páginas 172-185. En: J. Kacprzyk y M. Fedrizzi, Eds., *Multiperson Decision Making Using Fuzzy Sets and Possibility Theory*. Kluwer Academic Publishers, 1990.
- [9] L. Dombi. Fuzzy Logic and Soft Computing, *capítulo A General Framework for the Utility-Based and Outranking Methods*, páginas 202-208. World Scientific, 1995.
- [10] T. Tanino. Fuzzy preference orderings in group decision making. *Fuzzy Sets and Systems*, 12:117-131, 1984.
- [11] M. Roubens and Ph. Vincke. Preference modelling. Springer-Verlag, 1985.
- [12] C.Y. Yue, S.B. Yao, and P. Zhang. Rough approximation of a preference relation for stochastic multiattribute decision problems. *Lecture Notes in Artificial Intelligence*, 3613:1242-1245, 2005.
- [13] J. Kacprzyk. Group decision making with a fuzzy linguistic majority. *Fuzzy Sets and Systems*, 18(2):105-118, 1986.
- [14] F. Mata, L. Martínez, and E. Herrera-Viedma. An adaptative consensus support model for group decision making problems in a multigranular fuzzy linguistic context. *IEEE Transactions on Fuzzy Systems*, 17(2):279-290, 2009.
- [15] G. Campanella and R.A. Ribeiro. A framework for dynamic multiple-criteria decision making. *Decision Support Systems*, 52(1):52-60, 2011.
- [16] R.A. Ribeiro, T.C. Pais, and L.F. Simoes. Benefits of full reinforcement operators for spacecraft target landing. *Studies in Fuzziness and Soft Computing*, 257:353-367, 2010.
- [17] Ronald R. Yager and Alexander Rybalov. Uniform aggregation operators. *Fuzzy Sets and Systems*, 80(1):37-51, 1996.
- [18] H.J. Zimmermann and P. Zysno. Latent connectives in human decision making. *Fuzzy Sets and Systems*, 4(1):37-51, 1980.
- [19] Resolución 04169/1994 - Reglamento de funcionamiento del Consejo Superior de la UTEM.
- [20] L.I. Kuncheva and R. Krishnapuram. A fuzzy consensus aggregation operator. *Fuzzy Sets and Systems*, 79(3):347-356, 1995.
- [21] F.J. Cabrerizo, S. Alonso, I.J. Perez, and E. Herrera-Viedma. On consensus measures in fuzzy group decision making. *Lecture Notes in Computer Science*, 5285:86-97, 2008.
- [22] H. Legind. Efficient importance weighted aggregation between min and max. *9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002)*, 2002.
- [23] Fernández-Salido, J.M. y Murakami, S. Extending Yager's orness concept for the OWA aggregators to other mean operators. *Fuzzy Sets and Systems*, 139:515-542, 2003.
- [24] Dubois, D. y Prade, H. Fuzzy sets and systems: theory and applications. *Academic Press*. San Diego, 1980.
- [25] I. Palomares. Desarrollo de un Sistema Multi-Agente para automatizar Procesos de Consenso en Problemas de Toma de Decisión en Grupo. Jaén, 2010.
- [26] I. Palomares. Sistema Multiagente para modelar Procesos de Consenso en Toma de Decisión en Grupo a Gran Escala usando técnicas de Soft Computing. Jaén, 2014.
- [27] L. Escobar, P. Sánchez, L. Martínez. Modelo de Consenso basado en estructura de decisión dinámica. *ESTYLF XVII Congreso Español sobre Tecnologías y Lógica Fuzzy*. Zaragoza, España, Febrero 2014.
- [28] M. Espinilla, L. Martínez, S. Zapata C., Modelo Lingüístico de Toma de Decisiones Dinámicas Multicriterio con Información Heterogenea. *ESTYLF XVII Congreso Español sobre Tecnologías y Lógica Fuzzy*. Zaragoza, España, Febrero 2014.

# Prediction of Fatigue Crack Propagation in Bonded Joints Using Fracture Mechanics

Reza Hedayati, Meysam Jahanbakhshi

**Abstract--** Fracture Mechanics is used to predict debonding propagation in adhesive joint between aluminum and composite plates. Three types of loadings and two types of glass-epoxy composite sequences: [0/90]<sub>2s</sub> and [0/45/-45/90]<sub>s</sub> are considered for the composite plate and their results are compared. It was seen that generally the cases with stacking sequence of [0/45/-45/90]<sub>s</sub> have much shorter lives than cases with [0/90]<sub>2s</sub>. It was also seen that in cases with λ=0 the ends of the debonding front propagates forward more than its middle, while in cases with λ=0.5 or λ=1 it is vice versa. Moreover, regardless of value of λ, the difference between the debonding propagations of the ends and the middle of the debonding front is very close in cases λ=0.5 and λ=1. Another main conclusion was the non-dimensionalized debonding front profile is almost independent of sequence type or the applied load value.

**Keywords—** Adhesive; APDL; Debonding; Fatigue; Paris Law.

## I. INTRODUCTION

**A**DHESIVE bonding of aerospace components is a fabrication technique which, though over 70 years old, has increased markedly in popularity during the last two decades and is currently a focal point in many studies regarding aging aircraft [1]. In this work, Fracture Mechanics is implemented to predict debonding propagation in joint between aluminum and composite plates by means of 3D finite element analyses (Fig 1a). Three types of loadings: λ = 0, λ = 0.5 and λ = 1 and two types of glass-epoxy composite sequences: [0/90]<sub>2s</sub> and [0/45/-45/90]<sub>s</sub> are considered for the composite plate. Therefore 2\*3=6 cases are considered and their results are compared. Afterwards, the sequence [0/90]<sub>2s</sub> is called Sequence 1, and [0/45/-45/90]<sub>s</sub> is called Sequence 2. The durability, debonding face profile, and stress distribution will be compared between the six cases considered. A typical debonding face shape is shown in Fig. 1b.

In order to predict debonding propagation, the following equation called Paris Law is used:

$$\frac{da}{dN} = c(\Delta K_{eq})^m \quad (1)$$

where a is the debonding propagation, N is the number of cycles, K<sub>eq</sub> is the equivalent stress intensity factor. c and m are constants which have to be calculated using experimental results for each particular material. The equivalent stress intensity factor can be calculated using  $\Delta K_{eq} = \sqrt{K_I^2 + 2K_{II}^2}$ . In this study, the adherent chosen for bonding aluminum and composite plates is FM 300. The material properties of the FM300 adhesive are: E=2.73 GPa, σ<sub>y</sub> = 50 MPa. In [1], the constants have been given for the alternative Paris Law equation

$$\frac{da}{dN} = B(\Delta G)^d \quad (2)$$

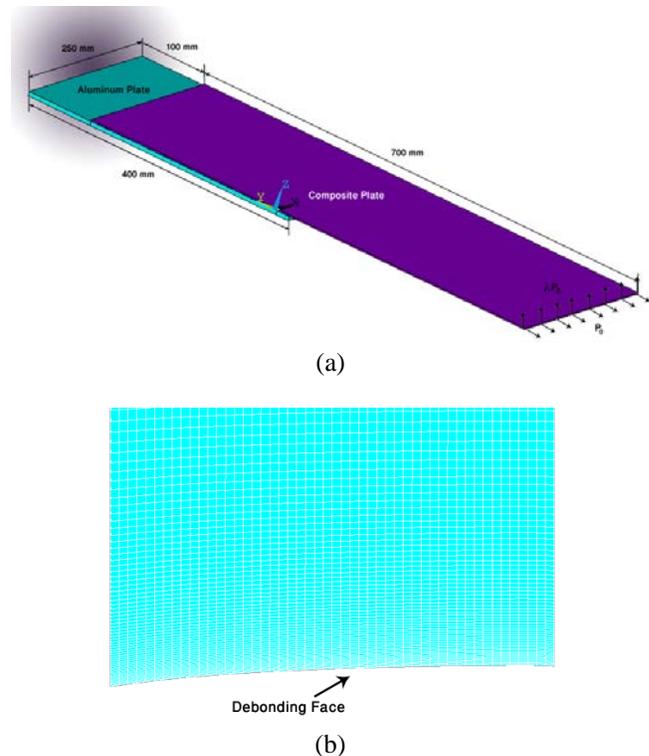


Fig. 1(a) The composite/Aluminum Joint Studied, (b) A typical debonding face shape

As b = 1.5\*10<sup>-15</sup>, and d=4.55 [2]. Using  $\Delta G = \frac{\Delta K^2}{E}$ , the constants for the initially introduced Paris Law (i.e. Eq. (1)) can be calculated by  $c = \frac{B}{E^d} \text{ and } m = 2d$ .

R. Hedayati is with the Young Researchers and Elite Club, Najaf Abad Branch, Islamic Azad University, Najaf Abad, Isfahan, Iran (corresponding author; e-mail: rezahedayati@gmail.com).

M. Jahanbakhshi is with the Young Researchers and Elite Club, Najaf Abad Branch, Islamic Azad University, Najaf Abad, Isfahan, Iran (e-mail: jahanbakhshimeysam@gmail.com).

## II. FINITE ELEMENT MODELING

In this project, a macro program is developed using ANSYS Parametric Design Language (APDL) to model debonding growth. At each step, the debonding face propagation, which is non-uniform, is calculated. Then the elements are completely cleared and a new model which consists of the updated debonding face is created and then meshed. This mesh deletion and creation is done at each propagation step in order to keep the accuracy of calculations well.

The major steps of the developed Macro program are as follows:

- (1) Define material properties of the model,
- (2) Define initial cycle increment  $\Delta N$  (usually about 1000 cycles). Also initially define cycle number  $N=0$ ,
- (3) Generate the geometry and mesh of the composite and aluminium plates and the adhesive,
- (4) Define the loading and constraints,
- (5) Perform the linear elastic solution,
- (6) Calculate the equivalent stress intensity factor ( $\Delta K_{eq}$ ) at each node located on the debonding face,
- (7) Calculate the debonding increment at each node located on the debonding face using Paris Law,
- (8) Calculate the new debonding front shape,
- (9) Sum up the old cycle number and the cycle increment ( $N_{New} = N_{Old} + \Delta N$ ),
- (10) If the maximum debonding increment is larger than 0.6 mm, then divide the cycle increment by 8,
- (11) If the maximum debonding increment is smaller than 0.3 mm, then multiply the cycle increment by 2,
- (12) If the debonding has reached the end of the adhesive or the mean shear stress is larger than yield shear stress of the adhesive, then stop the solution,
- (13) Delete the old geometry and mesh, and return to step (3).

TABLE I

MATERIAL PROPERTIES OF THE FM300 ADHESIVE AND ALUMINIUM 2024 [5]

| Property           | FM300                 | Aluminium 2024 |
|--------------------|-----------------------|----------------|
| Elasticity Modulus | 2.73 GPa              | 72 GPa         |
| Yield Stress       | 50 MPa                | 280 MPa        |
| Poisson's ratio    | 0.33                  | 0.27           |
| b (Paris constant) | $1.5 \times 10^{-15}$ | -              |
| d (Paris constant) | 4.5                   | -              |

The finite element model of the problem is shown in Fig. 2. For the composite plate 6000 8-noded SOLID46 elements, for the aluminum plate 22000 8-noded SOLID45 elements, and for the adhesive 8000 SOLID45 elements have been used. For the composite plate, the aluminium plate and

the adhesive, one, four and two elements through the thickness have been used. The elements at the two interfaces are glued. In other words, the composite and the adhesive share the same nodes at their interface. The same is true about the aluminium and the adhesive interface. This can be better seen in Fig. 2. Since the structure is symmetrical with respect to a plane perpendicular to X direction, only half of the model is created. The nodes located at the symmetry plane position are not allowed to move in X direction. A more complete information on the materials and method can be found in [3] and [4]. The material properties of the aluminium and the adhesive are listed in Table I.

## III. RESULTS AND DISCUSSION

### A. Stress Distribution

Running the code, it was seen that the debonding front gets two types of shapes for different loadings. For the cases with  $\lambda = 0$ , the debonding front is like a circle arc with its ends curved towards the positive Y direction (Fig. 3a), while for the cases with  $\lambda = 0.5$  and  $\lambda = 1$ , the debonding front is like a circle arc with its ends curved towards the negative Y direction (Fig. 3b). Fig. 3a shows the Von-Mises stress contour for the case  $\lambda = 0$  and sequence 1. For this case (Fig. 3), the applied load was chosen to be  $P_0 = 64 \text{ kN/m}$ . At the beginning (propagation of 10mm), the maximum stress intensity factor on the debonding face was  $1050 \text{ kPa}\sqrt{\text{m}}$ , while the minimum was  $938 \text{ kPa}\sqrt{\text{m}}$ . The maximum Von-Mises stress on the debonding front was 17MPa, and 18.7 MPa at the beginning ( $a=10\text{mm}$ ) and at the end ( $a=270\text{MPa}$ ), respectively. Therefore it can be concluded that the stress distribution and the debonding profile does not change a lot while propagation.

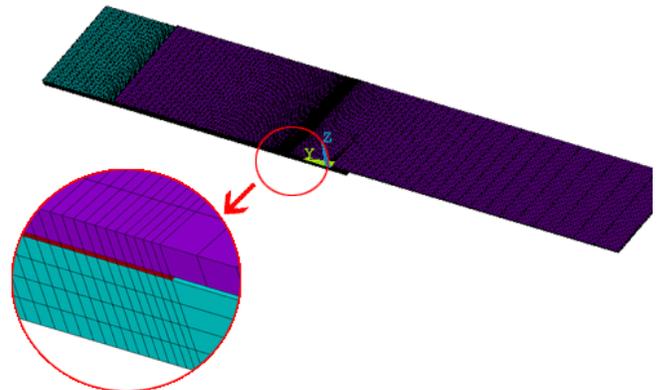


Fig. 2 Finite Element model of the Aluminium/Composite Joint

Fig. 4 shows the Von-Mises stress contour for the case  $\lambda = 1$  and sequence 1. For this case, the applied load was chosen to be  $P_0 = 0.8 \text{ kN/m}$ . The reason for choosing a much smaller load than the case with  $\lambda = 0$  is that in this case a load in Z direction is applied to the composite. Therefore the opening fracture mode has a very greater effect on the

adhesive. As a result if  $P_0$  is chosen to be large, then the composite debonds from aluminum immediately. As it can be seen from Fig.3b, unlike the case with  $\lambda = 0$ , for this case the maximum Von-Mises stress at the debonding face is 14.2 and 91.5MPa at the beginning and at the end which shows a huge increase.

For the case with  $\lambda = 0.5$ , the applied load was chosen to be  $P_0 = 1.04$  kN/m. The reason for choosing this load is to have the same force resultant for the cases  $\lambda = 0.5$  and  $\lambda = 1$ . Then it will be possible to compare their results at the same load.

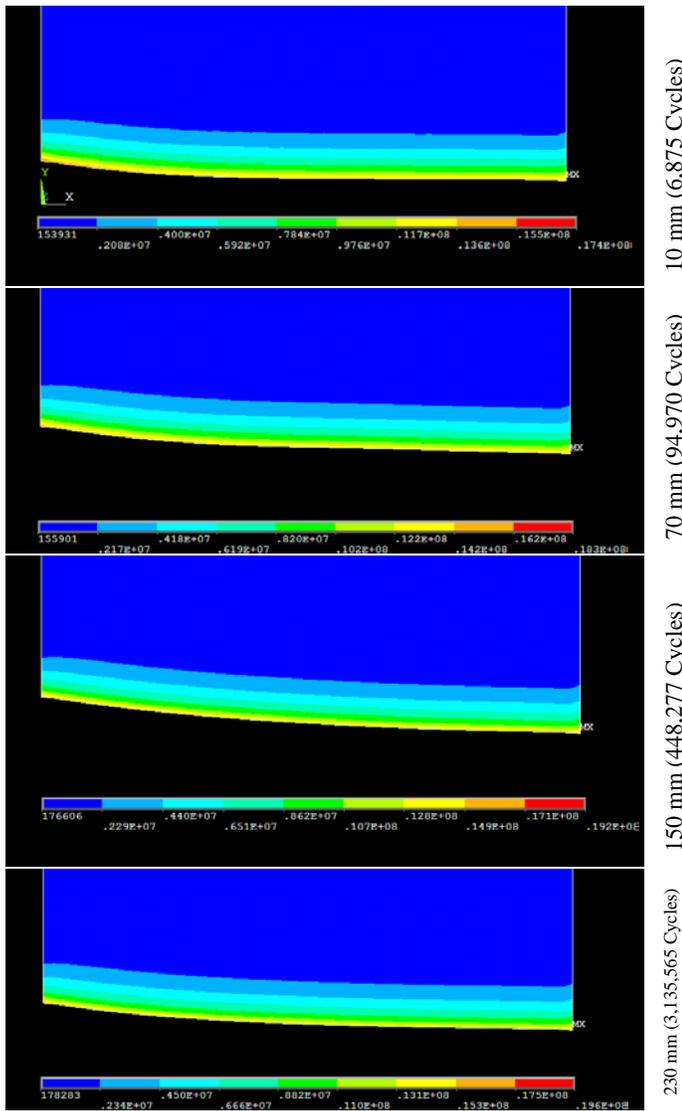


Fig. 3 Von-Mises stress contour for the case  $\lambda = 0$  and sequence 1

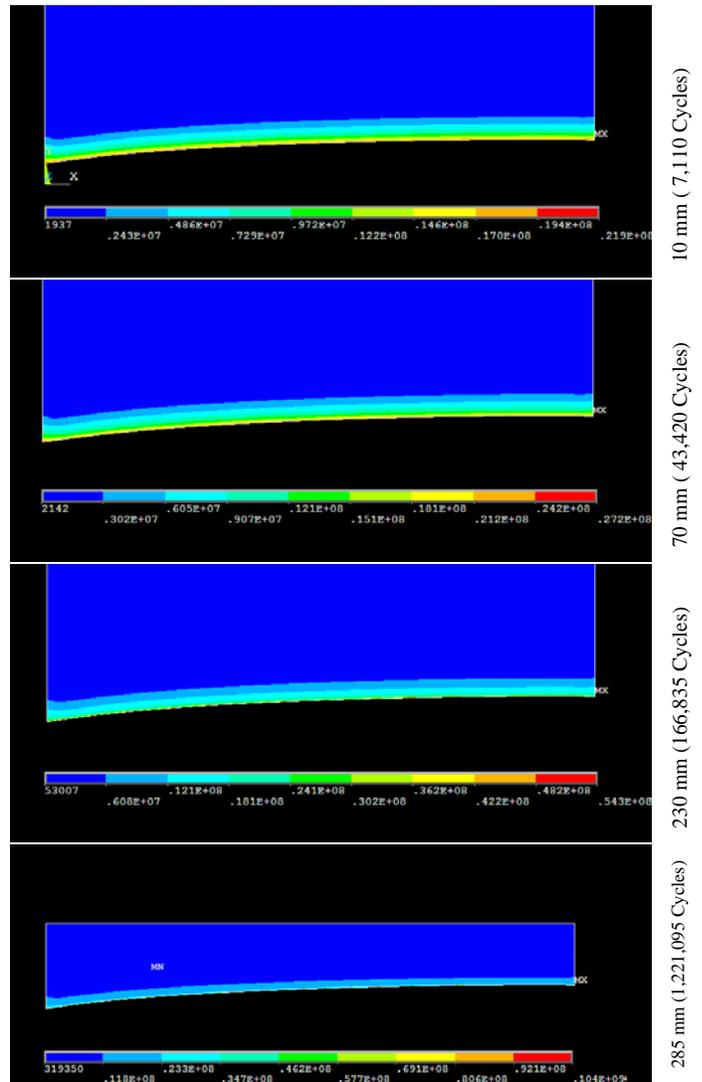


Fig. 4 Von-Mises stress contour for the case  $\lambda = 1$  and sequence 1

*B. Durability of the Structures*

The fatigue durability of the six cases is listed in Table II. As it can be seen from the table, the cases with stacking sequence 2 have much shorter lives generally. This is because using a composite having sequence 2 transfers more stress to the adhesive crack front in comparison with case with composite sequence 1, which causes an increase in  $K_I$  and  $K_{II}$ . Since for FM300 adhesive,  $K_{eq}$  has a power of about 9 in Paris Law equation, then increasing the stress at the debonding face increases the crack propagation speed very severely. This explanation is also true for why the cases with  $\lambda = 1$  have lower life cycles than the corresponding cases with  $\lambda = 0.5$ .

TABLE II  
FATIGUE DURABILITY OF THE SIX CASES

|            | $\lambda = 0$ | $\lambda = 0.5$ | $\lambda = 1$ |
|------------|---------------|-----------------|---------------|
| Sequence 1 | 198,333,972   | 65,985,225      | 1,221,095     |
| Sequence 2 | 4,964,955     | 8,216,448       | 275,618       |

### C. Debonding Face Profile

The difference between the debonding propagations of the ends and the middle of the debonding front is listed for the six cases in Table III. If one plots debonding face profile for different debonding propagations for cases with  $\lambda=0$  and  $\lambda=1$  and compare the curves, and also by comparing the results of Table III, one can see that:

- For both the sequence types, in cases with  $\lambda=0$  the ends of the debonding front propagates forward more than its middle part, while in cases with  $\lambda=0.5$  or  $\lambda=1$  the middle part of the debonding front moves forward more than its ends.
- For all values of  $\lambda$ , the difference between the debonding propagations of the ends and the middle of the debonding front of cases with composite sequence of 2 is higher than that for sequence 1. This can be more recognized when  $\lambda=0$ .
- For both the sequence types, the difference between the debonding propagations of the ends and the middle of the debonding front of the case with  $\lambda=0$  is higher than that in the corresponding case with  $\lambda=0.5$  or  $\lambda=1$ .
- Regardless of the sequence type, when  $\lambda=0$  the debonding face profile can be divided in three regions: (a) at the beginning of debonding propagation, the difference between the debonding propagations of the ends and the middle of the debonding front is small, (b) when the maximum propagation of the debonding front is higher than 50 mm, the difference between the debonding propagations of the ends and the middle of the debonding front gets larger and remains almost constant until near the end of propagation, and, (c) when the debonding front has reached near the end of adhesive film, the difference between the debonding propagations of the ends and the middle of the debonding front gets small again.
- If the non-dimensionalized debonding front profile is plotted for all the cases, it can be seen that the non-

dimensionalized debonding front profile is independent of sequence type or applied load value.

TABLE III  
DIFFERENCE BETWEEN THE DEBONDING PROPAGATION BETWEEN THE ENDS AND THE MIDDLE OF THE DEBONDING FRONT

|            | $\lambda = 0$ | $\lambda = 0.5$ | $\lambda = 1$ |
|------------|---------------|-----------------|---------------|
| Sequence 1 | 9 mm          | 7 mm            | 6.5 mm        |
| Sequence 2 | 28 mm         | 16 mm           | 15 mm         |

### IV. CONCLUSIONS

In this paper, Fracture Mechanics was used to predict debonding propagation in the adhesive joint between aluminium and composite plates. Three types of loadings:  $\lambda = 0$ ,  $\lambda = 0.5$  and  $\lambda = 1$  and two types of glass-epoxy composite sequences: [0/90]2s and [0/45/-45/90]s were considered for the composite plate. It was seen that generally the cases with stacking sequence [0/45/-45/90]s have much shorter lives than cases with [0/90]2s. About the debonding front profile, it was seen that for both the sequence types, in cases with  $\lambda=0$  the ends of the debonding front propagates forward more than its middle, while in cases with  $\lambda=0.5$  or  $\lambda=1$  the middle part of the debonding front moves forward more than its ends. It was also seen that regardless of  $\lambda$ , the difference between the debonding propagations of the ends and the middle of the debonding front is very close in cases  $\lambda=0.5$  and  $\lambda=1$ . Another main conclusion was the non-dimensionalized debonding front profile is almost independent of sequence type or the applied load value.

### REFERENCES

- [1] W.S. Johnson, L.M. Butkus, R.V. Valentin., "Applications of Fracture Mechanics to the Durability of Bonded Composite Joints", DOT/FAA/AR-97/56.
- [2] H. Hosseini-Toudeshky, B. Mohammadi. "Mixed-mode numerical and experimental fatigue crack growth analyses of thick aluminium panels repaired with composite patches". *Composite Structures*, vol. 91, pp. 1-8, 2009.
- [3] R. Hedayati, S. Ghorbani Khouzani, Meysam Jahanbakhshi. "Investigation of debonding propagation in aluminum/composite joints under fatigue loading." *Journal of Adhesion Science and Technology*, ahead-of-print pp. 1-15. 2014.
- [4] R. Hedayati, M. Jahanbakhshi, S. Ghorbani Khouzani. "Prediction of static crack propagation in adhesive joints." *Journal of Theoretical and Applied Mechanics*, vol. 52, no. 4, pp. 937-946, 2014.
- [5] R. Hedayati., S. Ziaei-Rad. "Foam-core effect on the integrity of tailplane leading edge during bird-strike event." *Journal of Aircraft*, vol. 48, no. 6 pp. 2080-2089, 2011.

# Plastic deformation and fracture processes in layered metal-graphene composites and polycrystalline graphene

Ilya A. Ovid'ko and Alexander G. Sheinerman

**Abstract**— We suggest theoretical models which describe fracture of polycrystalline graphene and competition between plastic deformation and fracture processes in metal-graphene layered composites. In considering polycrystalline graphene, we consider formation of cracks at grain boundaries (GBs) containing defects (partial disclinations and their dipoles) associated with experimentally observed structural irregularities of real GBs in graphene. Within the suggested model, we calculate the dependences of the critical stress for crack formation on the parameters of individual disclinations and their dipole configurations at GBs. We demonstrate that individual disclinations and their dipoles at GBs can be responsible for the experimentally observed (Huang *et al.* 2011 *Nature* **469** 389; Ruiz-Vargas *et al.* 2011 *Nano Lett.* **11** 2259) dramatic decrease of fracture strength of polycrystalline graphene compared to its pristine counterpart.

In consideration of metal-graphene layered composites, we examine the transfer of plastic deformation across a graphene interface and nanocrack formation initiated by stress fields of lattice dislocations stopped near a graphene interface. We reveal strength characteristics of metal-graphene layered composites as functions of their key structural parameters, including the metallic and graphene layer thicknesses, which are well consistent with the corresponding experimental data (Kim *et al.*, *Nature Commun.* **4** (2013) 2114). The results demonstrate that strong metal-graphene layered composites (against both fracture and macroscopic plastic flow) should contain monolayer graphene inclusions, and for such inclusions the processes of plastic deformation and interface fracture compete and can occur concurrently.

**Keywords**—graphene; grain boundaries; composites; defects; cracks

The work was supported by St. Petersburg State University research grant 6.37.671.2013 and the Russian Ministry of Education and Science (Grants 14.B25.31.0017 and MD-2893.2015.1).

I.A. Ovid'ko is with Research Laboratory for Mechanics of New Nanomaterials, St. Petersburg State Polytechnical University, St. Petersburg 195251, Russia, Department of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg 198504, Russia, and Institute of Problems in Mechanical Engineering, Russian Academy of Sciences, St. Petersburg 199178, Russia (email: [ovidko@gmail.com](mailto:ovidko@gmail.com)).

A.G. Sheinerman is with Research Laboratory for Mechanics of New Nanomaterials, St. Petersburg State Polytechnical University, St. Petersburg 195251, Russia, Department of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg 198504, Russia, and Institute of Problems in Mechanical Engineering, Russian Academy of Sciences, St. Petersburg 199178, Russia (phone: +7812 321 4764; fax: +7812 321 4771; email: [asheinerman@gmail.com](mailto:asheinerman@gmail.com)).

## I. INTRODUCTION

Graphene – a single carbon atomic sheet with the hexagonal  $sp^2$  covalently bonded crystal structure – with its outstanding mechanical, transport and thermal properties represents the subject of rapidly growing research efforts in applied physics and materials science [1–6]. Of crucial importance from both fundamental and applied viewpoints is the unique behaviour of graphene under mechanical load. In particular, Lee with co-workers have experimentally demonstrated that pristine graphene exhibits the highest ever measured strength of  $\approx 130$  GPa (Ref. [7]). At the same time, following experimental examinations [8,9], the strength characteristics of graphene sheets containing GBs dramatically degrade compared to the superior strength ( $\approx 130$  GPa) of their pristine counterparts. These experimental data motivate large interest in understanding the physical mechanisms of fracture in graphene and their sensitivity to the presence of defects. Therefore, in the second section of this paper we suggest a theoretical model describing crack generation at elementary irregularities of the GB structure, namely, those associated with partial disclinations and their dipoles.

Also, since graphene monolayer sheets and multilayer nanoplatelets are specified by superior values of strength and elastic moduli, they are very good candidates for the use as reinforcing structural elements in polymer-, ceramic- and metal-matrix composites. For instance, recently, Kim with co-workers [10] have synthesized Cu- and Ni-graphene nanolayered composites exhibiting extremely high strength characteristics (with the flow stress at 5% strain of 1.5 GPa for Cu-graphene composites and 4.0 GPa for Ni-graphene composites). Also, Kim with co-workers [10] experimentally revealed that the above flow stress increases with diminishing the metal layer thickness.

The dominant physical mechanism responsible for superior strength of metal-graphene nanolayered composites is attributed to the role of graphene interfaces as obstacles for lattice dislocation glide [10]. In the case under consideration, it is logical to think that the plastic deformation and fracture processes controlling the flow stress/strength of a metal-graphene layered composite are the transfer of plastic

deformation across a graphene interface and the nanocrack formation initiated by stress fields of lattice dislocation stopped near a graphene interface. In the third section we describe the strength-controlling processes (transfer of plastic deformation and nanocrack generation) in metal-graphene layered composites and reveal the dependence of their strength characteristics on the metallic and graphene layer thicknesses.

II. CRACK NUCLEATION AT PARTIAL DISCLINATIONS AT GRAIN BOUNDARIES IN GRAPHENE

In general, GBs in 2D graphene are line defects separating graphene grains (crystallites/domains) whose crystal lattices are tilted by a non-zero angle  $\theta$  relative to each other [11]. The angle  $\theta$  serves as the main geometric parameter of a GB and is called the GB misorientation. According to experimental data, computer simulations and theoretical models, low- and high-angle GBs in graphene hexagonal lattices are represented as walls of edge dislocations or, in other terms, pentagon-heptagon pairs [8]. The geometry of the individual geometries of pentagon-heptagon pairs and their spatial arrangement in the corresponding wall configuration determines GB misorientation.

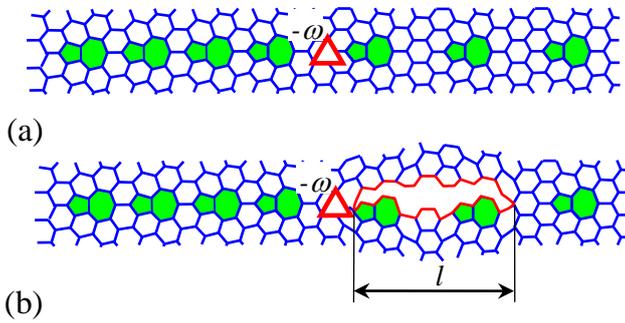


Fig. 1. A partial disclination at grain boundary in graphene bi-crystal. (a) Graphene bi-crystal without cracks. (b) Graphene bi-crystal containing a nanocrack

Within our model, we consider GB defects associated with elementary changes in the GB misorientation and their effects on crack generation at GBs. Such defects are called partial disclinations in a 2D graphene sheet and represent the points where the GB misorientation changes in a step-like manner (and so does the GB dislocation arrangement; see figure 1(a)), so that the jump of misorientation represents the disclination strength  $\omega$ ; see Ref. [11]. With the hexagonal geometry of the graphene crystal lattice, the strengths  $\omega$  of such partial disclinations can be arbitrary in the range:  $-60^\circ < \omega < 60^\circ$ . Partial GB disclinations create stresses that can initiate nanocracks in graphene (figure 1(b)).

First, consider crack generation in a flat graphene sheet with a line GB containing a single partial disclination of the strength  $-\omega$  (figure 1(b)). Within our model, the flat graphene sheet has a circular shape specified by the radius  $R$ ,

and the partial disclination is located at its center (figure 2). (The radius  $R$  plays the role of the screening length for the stresses created by the disclination.) Consider the situation where the flat graphene sheet is under a tensile mechanical load  $\sigma_0$  whose direction is normal to the GB line (figure 2). The disclination creates local stresses which, in superposition with the external load, can initiate nanocrack formation. Within our model, the nanocrack nucleates and grows along a GB in the region where the tensile stresses exerted on the crack surfaces by the disclination and the applied load are highest (see figure 1(b)).

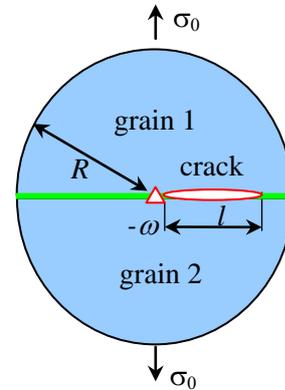


Fig. 2. Nanocrack generation at a grain boundary disclination (triangle) located at center of mechanically loaded circular graphene sheet

In order to calculate the conditions for nanocrack growth in graphene, we use the energy criterion [12] suggesting that a crack is favored to grow if the release of the strain energy in the course of crack advance is larger than the effective surface energy of the crack surfaces. With this criterion, using the expressions for the stress field of a wedge disclination in an infinite medium (and modifying these for the case of the plane stress state), one obtains the following condition for catastrophic crack growth:  $\sigma > \sigma_c$ , where

$$\sigma_c = (8D(2\gamma - \gamma_b) / l_0)^{1/2} - D\omega(\ln(4R / l_0) - 2). \quad (1)$$

In formula (1),  $l_0$  represents the maximum crack length to which the crack can grow through thermal fluctuations,  $R$  is the screening length of the disclination stress field (in our case,  $R$  is the graphene circle radius; see figure 4),  $D = E / (4\pi)$ ,  $E$  is the Young modulus of graphene,  $\gamma$  is the specific surface energy of graphene edges (say, crack edges in graphene), and  $\gamma_b$  is the specific GB energy in graphene. In derivation of formula (1), it is assumed that  $l \ll R$ . The quantities  $\gamma$  and  $\gamma_b$  have the meaning of the energy per unit area, that is, the energy of the surface of a graphene edge (or graphene GB energy, respectively) per length of the graphene edge (GB, respectively) divided by the distance (0.34 nm) between the graphene sheets in graphite.

In order to calculate the critical stress  $\sigma_c$  for a GB crack, we use the following typical values of graphene

characteristics:  $l_0 = 0.72$  nm,  $E = 1000$  GPa (Ref. [12]),  $\gamma = 10.3$  J/m<sup>2</sup> [14], and  $\gamma_b = 3$  J/m<sup>2</sup> [15]. The dependences of the critical stress  $\sigma_c$  on the disclination strength  $\omega$  are calculated and presented in figure 3 for various values of the screening length (the radius of the circular graphene sheet)  $R$ . As it follows from figure 3, when  $\omega$  increases,  $\sigma_c$  decreases from 125 GPa at  $\omega = 0$  down to zero at certain values of  $\omega$  (10 to 20 degrees, depending on the value of the screening length  $R$ ). The calculated low values of the critical stress  $\sigma_c$  for intergranular fracture are consistent with the experimentally documented [9] values (35 GPa or lower) of fracture stresses specifying polycrystalline graphene specimens. As a corollary, individual partial disclinations at GBs can serve as critical defects responsible for experimentally documented [8,9] dramatic decrease in strength of polycrystalline graphene, as compared to its pristine counterpart.

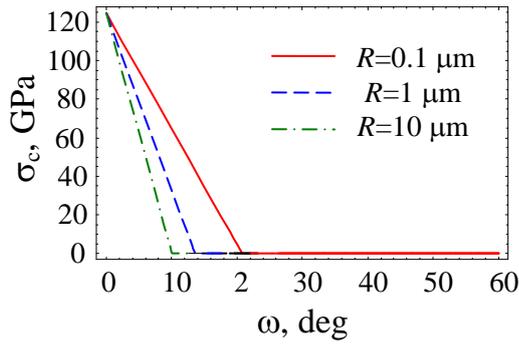


Fig. 3. Dependences of the ultimate stress  $\sigma_c$  on disclination strength  $\omega$  for various values of the screening length  $R$

Besides individual disclinations, we have analyzed the effect of disclination dipoles (representing two opposite-sign disclinations) on the critical stress for catastrophic crack growth. The analysis has demonstrated that the critical stress  $\sigma_c^{dip}$  for crack generation at a disclination dipole and its catastrophic growth decreases with increasing the disclination strength and the distance between the dipole disclinations. For large enough values of these parameters, the critical stress  $\sigma_c^{dip}$  is smaller than 35 GPa. Thus, similar to individual partial disclinations at GBs, their dipoles can be responsible for the experimentally documented [8,9] dramatic decrease in the strength of polycrystalline graphene.

Similar to grain boundaries in graphene, graphene interfaces in metal-graphene layered composites can serve as weak elements that can decrease the strength of such composites. Therefore, nanocrack generation and transfer of plastic flow across graphene interfaces in metal-graphene layered composites will be examined in the next section.

### III. COMPETITION BETWEEN NANOCRACK GENERATION AND TRANSFER OF PLASTIC FLOW ACROSS GRAPHENE INTERFACES IN METAL-GRAPHENE LAYERED COMPOSITES. GENERAL ASPECTS

Consider a layered composite solid consisting of repeat metallic layers and graphene interfaces. Let the solid be under the action of a shear stress  $\tau$ . Consider an ensemble of  $N$  rectangular glide dislocation loops with identical Burgers vectors  $\mathbf{b} = b\mathbf{e}_y$  (where  $\mathbf{e}_y$  is the unit vector directed along the  $y$ -axis) formed due to the action of a Frank-Read source and stopped near a platelike impenetrable graphene layer (figure 4). The action of the applied stress  $\tau$  and the stress field created by the ensemble of dislocation loops (stopped in the plastically deformed layer located to the left side of the graphene inclusion/interface) can induce homogeneous generation of a new rectangular glide dislocation loop with the Burgers vector  $\mathbf{b}$  in the neighboring metallic layer located to the right side of the graphene layer (figure 4).

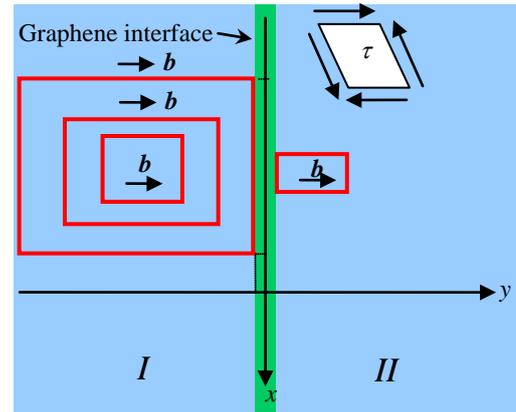


Fig. 4. Transfer of plastic deformation across a graphene interface. Transfer of plastic flow occurs through formation a rectangular glide dislocation loop in a metal layer  $I$  under the superposition of the applied shear stress  $\tau$  and the stress field of an ensemble of rectangular glide dislocation loops located in the neighboring metal layer  $II$

Within this model, we have calculated the critical shear stress  $\tau = \tau_{pl}$  for barrier-free generation of a new rectangular glide dislocation loop in Ni-graphene layered composite. The dependences of the critical shear stress  $\tau_{pl}$  on the parameter  $\lambda$  characterizing the thickness of Ni layers in the Ni-graphene layered composite are plotted in figure 5, for various values of the graphene interface thickness  $h$ . It is seen in figure 5 that the critical stress  $\tau_{pl}$  decreases with an increase in the metal layer thickness  $\lambda$  and/or a decrease in the graphene layer thickness  $h$ . For a given value of  $\lambda$ , the stress  $\tau_{pl}$  is minimum in the case of a monolayer graphene interface having the thickness of  $h \approx 0.3$  nm.

Thus, plastic deformation can be transferred through graphene interfaces at high local stresses created by

dislocation pileups. At the same time, the high stresses concentrated near the head of a dislocation pileup can induce the formation of a nanocrack: either in the metallic layer, at the angle  $\alpha$  to the normal to the layer boundary (figure 6(a)), or at the metal-graphene interface (figure 6(b)), or within the graphene interface (if this interface represents a multilayer graphene sheet; figure 6(c)).

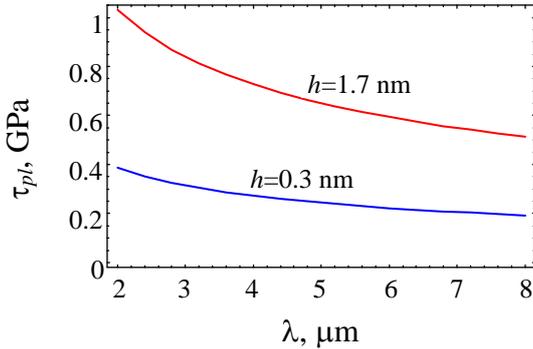


Fig. 5. The critical shear stress  $\tau_{pl}$  for the barrier-free formation of a dislocation loop at a graphene interface vs the thickness  $\lambda$  of Ni layers in a layered Ni-graphene composite, for different values of the graphene layer thickness  $h$

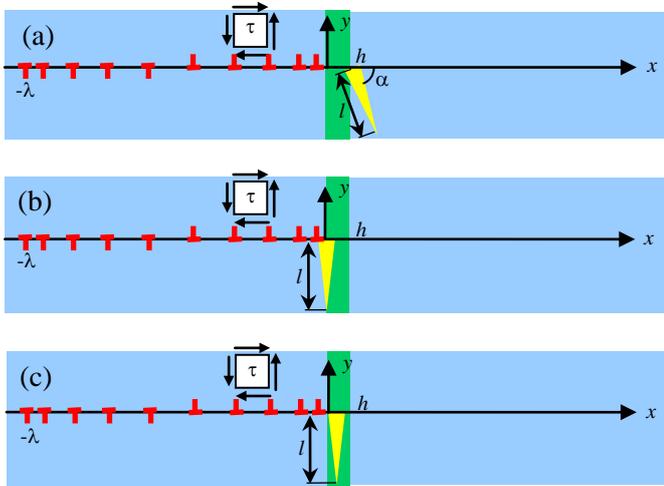


Fig. 6. Generation of nanocracks at a graphene interface in a metal matrix in the stress field of a double dislocation pileup and the applied shear stress  $\tau$ . (a) Nanocrack forms in the matrix. (b) Nanocrack forms at the matrix-graphene interface. (c) Nanocrack forms inside the graphene inclusion

We have calculated the condition for the generation of a nanocrack at the head of the dislocation pileup formed under the applied shear stress  $\tau$  (figure 6). This condition has the form  $\tau > \tau_{fr}$ , where  $\tau_{fr}$  is the critical stress for nanocrack generation. The dependences of the critical stress  $\tau_{fr}$  for nanocrack generation in the Ni-graphene layered composite on the parameter  $\lambda$  are shown in figure 7, for the cases

illustrated in figures 6(a), 6(b) and 6(c). The two upper curves in figure 7 correspond to the formation of a nanocrack in the Ni layer (figure 6(a)), for  $\alpha = 70^\circ$ ,  $h = 1.7$  nm and 0.3 nm (curves 1 and 2, respectively). Curve 3 corresponds to the formation of a nanocrack at the Ni-graphene interface (figure 6(b)). Curve 4 corresponds to the formation of a nanocrack within the multilayer graphene platelet at one interatomic distance from the Ni-graphene interface terminating the dislocation pileup (figure 6(c)).

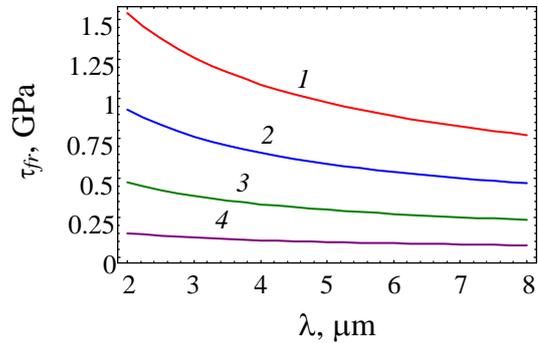


Fig. 7. Dependences of the critical stress  $\tau_{fr}$  for nanocrack generation on the thickness  $\lambda$  of Ni layers in a layered Ni-graphene composite, for nanocrack in Ni (curves 1 and 2), Ni-graphene interface (curve 3) and graphene platelet (curve 4), with  $h = 1.7$  nm (curve 1) and 0.3 nm (curve 2). For curves 3 and 4, graphene inclusion thickness  $h$  is arbitrary

As it is seen in figure 7, the critical stress  $\tau_{fr}$  is the lowest for the case of nanocrack formation inside a multilayer graphene interface. At the same time, figure 7 demonstrates that if the graphene interface represents a monolayer graphene sheet, the nanocrack is the easiest to form along the Ni-graphene interface. Thus, the formation of nanocracks in metal-graphene layered composites is least likely, if the graphene interfaces represent monolayer graphene sheets and the metal layers adhere well with graphene.

Figure 8 plots the typical dependences of  $\tau_{pl}$  and  $\tau_{fr}$  on  $\lambda$  for layered Ni-graphene composites with monolayer and multilayer graphene sheets. Here the critical stress  $\tau_{fr}$  corresponds to the easiest way of crack formation, that is, at the Ni-graphene interface (figure 6(b)) (for monolayer graphene sheets) or within the multilayer graphene sheet at one interatomic distance from the Ni-graphene interface (for multilayer graphene sheets). Figure 8 shows that the formation of a nanocrack in a layered Ni-graphene composite with multilayer graphene interfaces occurs at much smaller stresses that the transfer of plastic deformation through the graphene interface by means of the formation of a new dislocation loop. At the same time, in a Ni-graphene nanolayered composite with monolayer graphene interfaces, for any values of  $\lambda$ , the critical stress  $\tau_{fr}$  for the formation of an interface nanocrack is close to the critical stress  $\tau_{pl}$  for the formation of a new

## IV. CONCLUSIONS

dislocation loop. Therefore, in the case of monolayer graphene interfaces in Ni-graphene nanolayered composite, the processes of plastic deformation and interface fracture compete and can occur simultaneously.

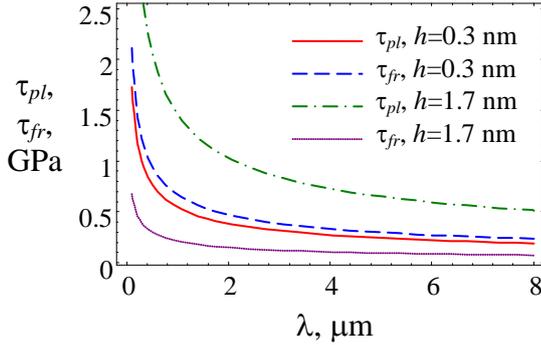


Fig. 8. Dependences of the critical stresses  $\tau_{pl}$  and  $\tau_{fr}$  on the thickness  $\lambda$  of the Ni layer in nanolayered Ni-graphene composites containing monolayer ( $h=0.3$  nm) or multilayer ( $h=1.7$  nm) graphene sheets

Figure 8 also allows one to make a rough comparison of experimental and calculated values of the strength characterizing Ni-graphene nanolayered composites with monolayer graphene interfaces. For Ni layer thickness of  $\lambda=100$  nm and the compressive load normal to the Ni-graphene interfaces, experimental data [10] demonstrate the transition from very strong strengthening (characterized by a very small increase in plastic deformation with increasing the applied load) to moderate strengthening (characterized by pronounced plastic flow) at the value of the applied load around 3 GPa. This value of the applied load corresponds to the minimum (critical) shear stress in planes incident to the interfaces of 1.5 GPa, which is consistent with the theoretically revealed value of  $\tau_{pl} \approx 1.75$  GPa, for  $\lambda=100$  nm (see figure 8). This enables us to assume that at the load below 3 GPa, small plastic deformation occurs in individual metal layers, while at load above 3 GPa the transfer of plastic deformation across graphene interfaces initiates pronounced plastic flow in Ni-graphene layered composites.

Let us define the critical stress  $\tau_m$  as the minimum stress at which the layered Ni-graphene composite can either deform with onset of macroscopic plastic deformation or fracture. With the dominant mechanism of plastic flow assumed to be the transfer of plastic deformation from one metal layer to another one across a graphene interface, we define  $\tau_m$  as  $\tau_m = \min\{\tau_{pl}, \tau_{fr}\}$ . Figure 8 convincingly demonstrates that the critical stress  $\tau_m$  for Ni-graphene layered composites containing multilayer graphene inclusions is much lower than that for the composites containing monolayer graphene interfaces. This means that strong Ni-graphene layered composites (against both fracture and macroscopic plastic flow) should contain monolayer graphene inclusions.

To summarize, we have theoretically examined fracture of polycrystalline graphene and competition between plastic deformation and fracture processes in metal-graphene layered composites. In considering polycrystalline graphene, we have demonstrated that individual disclinations and their dipoles at GBs can be responsible for the experimentally observed [8,9] dramatic decrease of fracture strength of polycrystalline graphene compared to its pristine counterpart.

In consideration of metal-graphene layered composites, we have demonstrated that plastic deformation and fracture processes in metal-graphene layered composites are in competition and crucially affects the ultrahigh strength of these composites. The results have shown that the formation of a nanocrack in a layered Ni-graphene composite with multilayer graphene inclusions occurs at much smaller stresses than the transfer of plastic deformation through the graphene inclusion/interface by means of the formation of a new dislocation loop. Also, the critical fracture stress  $\tau_{fr}$  for Ni-graphene nanolayered composite with multilayer graphene interfaces is always lower than the critical stresses  $\tau_{pl}$  and  $\tau_{fr}$  characterizing Ni-graphene nanolayered composites containing monolayer graphene interfaces, for the same values of the metal layer thickness  $\lambda$ . Therefore, Ni-graphene nanolayered composites containing monolayer graphene interfaces are specified by higher strength than their counterparts with multilayer graphene interfaces. At the same time, in a Ni-graphene nanolayered composite with monolayer graphene interfaces, for any values of the metal layer thickness  $\lambda$ , the critical stress  $\tau_{fr}$  for the formation of an interface nanocrack is close to the critical stress  $\tau_{pl}$  for the transfer of plastic deformation across a graphene interface. As a corollary, in the case of monolayer graphene interfaces in Ni-graphene nanolayered composite, the processes of plastic deformation and interface fracture compete and can occur concurrently.

## REFERENCES

- [1] A. K. Geim and K. S. Novoselov, "The rise of graphene," *Nature Mater.*, vol. 6, pp. 183–191, 2007.
- [2] A.K. Geim, "Graphene: status and prospects," *Science*, vol. 324, pp. 1530–1534, 2009.
- [3] A. H. Castro Nero, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, "The Electronic Properties of Graphene", *Rev. Mod. Phys.*, vol. 81, pp. 109–162, 2009.
- [4] A. A. Balandin, "Thermal properties of hraphene and nanostructured carbon materials," *Nature Mater.*, vol. 10, pp. 569–581, 2011.
- [5] I. A. Ovid'ko, "Mechanical properties of graphene," *Rev. Adv. Mater. Sci.*, vol. 34, pp. 1–11, 2013.
- [6] T. Yamada, J. Kim, M. Ishihara, and M. Hasegawa, "Low-temperature graphene synthesis using microwave plasma CVD," *J. Phys. D*, vol. 46, art. 063001, 2013.
- [7] C. Lee, X. Wei, J. W. Kysar, and J. Hone, "Measurement of the elastic properties and intrinsic strength of monolayer graphene," *Science*, vol. 321, pp. 385–388, 2008.

- [8] P. Y. Huang, C. S. Ruiz-Vargas, A. M. van der Zande, W. S. Whitney, M. P. Levendorf, J. W. Kevek, S. Garg, J. S. Alden, C. J. Hustedt, Y. Zhu, J. Park, P. L. McEuen, and D. A. Muller, "Grains and grain boundaries in single-layer graphene atomic patchwork quilts," *Nature*, vol. 469, pp. 389–392, 2011.
- [9] C. S. Ruiz-Vargas, H. L. Zhuang, P. Y. Huang, A. M. van der Zande, S. Garg, P. L. McEuen, D. A. Muller, R. C. Hennig, and J. Park, "Softened elastic response and unzipping in CVD graphene membranes," *Nano Lett.*, vol. 11, pp. 2259–2263, 2011.
- [10] Y. Kim, J. Lee, M. S. Yeom, J. W. Shin, H. Kim, Y. Cui, J.W. Kysar, J. Hone, Y. Jung, S. Jeon, S. M. Yan, "Strengthening effect of single-atomic-layer graphene in metal-graphene nanolayered composites," *Nature Commun.*, vol. 4, p. 2114, 2013.
- [11] I. A. Ovid'ko, "Review on grain boundaries in graphene. Curved nano- and polycrystalline graphene structures as new carbon allotropes," *Rev. Adv. Mater. Sci.*, vol. 30, pp. 201–224, 2012.
- [12] G. R. Irwin, "Analysis of stresses and strains near the end of crack traversing a plate," *J. Appl. Mech.*, vol. 24, pp. 361–364, 1957.
- [13] S. Arghavan and A. V. Singh, "Effects of van der Waals interactions on the nonlinear vibration of multi-layered graphene sheets," *J. Phys. D*, vol. 45, art. 455305, 2012.
- [14] K. Kim, V. I. Artyukhov, W. Regan, Y. Liu, M. F. Crommie, B. I. Yakobson, A. Zettl, "Ripping graphene: preferred directions," *Nano Lett.*, vol. 12, pp. 293–297, 2012.
- [15] A. Cao and Y. Yuan, "Atomistic study on the strength of symmetric tilt grain boundaries in graphene," *Appl. Phys. Lett.*, vol. 100, art. 211912, 2012.

# Complex Social Network Interactions in Coupled Socio-Ecological System: Multiple Regime Shifts and Early Warning Detection

Hendrik Santoso Sugiarto<sup>1,2</sup>, Lock Yue Chew<sup>1,2</sup>, Ning Ning Chung<sup>3</sup> and Choy Heng Lai<sup>3</sup>

<sup>1</sup>Division of Physics and Applied Physics, Nanyang Technological University, Singapore 637371

<sup>2</sup>Complexity Institute, Nanyang Technological University, Singapore 637723

<sup>3</sup>Department of Physics, National University of Singapore, Singapore 117542

**Abstract** - We investigate the effects of complex social network interactions on social regime shifts within a coupled socio-ecological system. We observe the occurrence of hysteresis between the cooperative and defective regime as we vary the resource inflow within the system. As we adjust the social network properties such as degree and topology, we notice a change in the width of the hysteresis curve. This result signifies the intimate connection between the underlying structure of the social interactions and the resiliency of the coupled socio-ecological system. In particular, we uncover a new feature of multiple regime shifts within the hysteresis curve as we introduce community structures into the complex social interactions, indicating that the presence of sub-structures in the interactions can break up the collapse or revival of a full regime shifts into multiple smaller regime shifts. Furthermore, we highlight the possibilities of making accurate early warning detections on the occurrence of regime shifts through both temporal and spatial indicators. We show that spatial indicators are more robust to changes in the degree of these social network interactions.

## 1. Introduction

In recent years, the interaction between humans and their environment has become intense and inevitable. Our future is threatened by the prospect of resource scarcity<sup>1,2</sup> and massive climate change<sup>3,4</sup>, with humans being the source of the significant deterioration of waters and its hydrologies<sup>5</sup>, forests<sup>6</sup>, as well as biodiversity<sup>7</sup>. Our ecosystem has experienced sudden, abrupt collapse and long lasting alteration to its structure. Human activities such as industrialization and exploitation are responsible for such collapse of ecosystems<sup>8</sup>. This occurrence is known in the literature as regime shifts. Regime shifts imply the existence of multiple stable states within the system<sup>9,10</sup>. It happens when gradual alteration of underlying parameters triggers an unexpected transition near a critical point from one stable regime to a new stable regime<sup>11</sup>. Multiple stable states also indicate that the state of the system depends not merely on its variables and parameters but also on the history of the system. This path dependency is known as hysteresis, which usually arises via a change of certain driving parameter. Note that such a system cannot be reversed to its original domain by merely returning the parameter to its previous value. This property of irreversibility within hysteresis makes regime shift

catastrophic, in the sense that one cannot return the situation back to its normal state by a simple retraction<sup>9</sup>.

In the case of socio-ecological regime shift, catastrophic transition can happen from a failure of cooperation. In real social interactions, the dynamics of cooperation is closely associated with the structure of social interaction. Thus, network properties will affect the local interaction among individuals which affect the multiple stabilities in the system. This in turn affects the bifurcation characteristics and the position of the tipping points of the system. Previous research has connected the relation between network properties and critical transition since the connectivity of network structure is usually associated with a resistance to change which impacts the critical transition between regimes<sup>12</sup>. Therefore, we intend to understand the robustness of different kinds of networks, for instance, what kind of topology would cause the overall systems response to be gradual or catastrophic. This knowledge would give us insights into specific social structures that are less vulnerable to collapse or exhibit the effects of hysteresis. In the context of social network, a vertex corresponds to an individual whereas the edges represent the social interactions. Many model employs an underlying network structure to depict social interactions to improve the reality of their system<sup>13-16</sup>. In real life scenario, social network structure often differs from one society to another. Some society also exhibit community structure in which individuals often interact closely within their own community<sup>17,18</sup>.

With the presence of risk in the sudden and persistent collapse of socio-ecological system, it will be very useful if we are able to anticipate regime shifts before the transition occurs<sup>12,19,20</sup>. For complex systems, a lack of detailed information makes it difficult to determine the exact position of the tipping point. To circumvent this difficulty, there is a rapid growth in the study of early warning signals of critical transition based on the generic behaviour in the vicinity of regime shifts<sup>21-28</sup>. Early warning signals can be used as an indicator of a pending regime shift. It gives us enough lead time to pre-empt the regime shift or to start evacuation procedure if the regime shift is unavoidable<sup>12,29</sup>. However, early warning signals cannot predict the future transition accurately all the time<sup>30-32</sup>. There are possibility of

false positive (when the early warning signal indicates an approaching transition but turns out to be false detection) and false negative (when the early warning signal failed to predict the approaching transition). In this paper, we will explore the applicability of conventional early warning signals such as autocorrelation and standard deviation to detect the occurrence of regime shifts accurately.

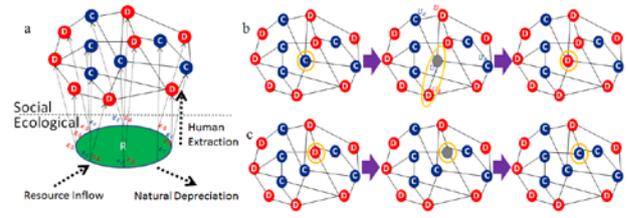
## 2. Model

A specific example of a simple coupled socio-ecological model is the Common Pool Dilemma. This problem is interestingly described by Hardin as a tragedy of the commons<sup>33</sup>. In this problem, the maintenance of the ecological resource requires cooperative behaviour among related individuals who have equal access to the common resource. The collapse of cooperation is inevitable if everyone is rational and selfish since every selfish act to maximize individual profits will lead to a depletion of resources which in turn destroys the economic viability of the whole system. Common pool dilemma is very relevant to our situation today because the increasing competition for the common ecological resources is the main culprit that damages our natural ecosystem.

In many common pool models, production is described by using the Cobb-Douglas function with decreasing returns, i.e.  $F = \gamma E^\alpha R^\beta$ , where  $E$  is the total effort and  $R$  the resource available<sup>34-36</sup>. The total payoff is then described by subtracting the opportunity cost from the production function:  $\pi_c = \frac{e_c}{E} F - we_c$  for co-operator, and  $\pi_d = \frac{e_d}{E} F - we_d$  for defector. Previously, Tavoni et al proposed an ostracism mechanism to maintain the cooperation among resource users<sup>36</sup>. This model is called the TSL model, which is formulated in the form of non-linear dynamical equations consisting of two main components: the social dynamics and the ecological dynamics. In this paper we retain all important features of the TSL model. TSL model employs an equity driven ostracism mechanism to maintain the cooperation level, which leads to the following utility:  $U_d(n_c) = \pi_d - O(n_c) \frac{\pi_d - \pi_c}{\pi_d}$  for defector and  $U_c = \pi_c$  for co-operator.  $O(n_c) = he^{te^{gn_c}}$  is the ostracism function with the parameter  $h$  representing the ostracism strength, and the parameters  $t$  and  $g$  govern the shape and effective threshold of the ostracism function. The rate of change of the available ecological resource is made up of 3 components: linear resource inflow, natural depreciation and human extraction:  $\frac{\Delta R}{\Delta t} = c - d \left[ \frac{R}{R_{max}} \right]^2 - qER$ . The central assumption of this model is a well-mixed social interaction, which makes the ostracism mechanism effective against defectors. To make it more realistic, we have modified the model by adding social network to constraint social interaction among users. We have also incorporated discrete updating so that the social and ecological variables are evaluated at every time step. The schematic image of this model is shown in figure 1a below.

The parameters used in the simulation are  $\alpha = 0.6$ ,  $\beta = 0.2$ ,  $\gamma = 10$ ,  $q = 1$ ,  $d = 50$ ,  $R_{max} = 200$ ,  $w = 15$ ,  $h = 0.34$ .

In this paper, all individuals interact locally with their adjacent neighbours in a specific social network. The social interaction here involves ostracism as social sanction with utility comparison. Moreover, we shall consider social interaction based on the Erdos-Renyi network topology, scale-free network topology, and also network with community structure. For the updating mechanism, we use asynchronous pairwise comparison such that at each time step a random player updates his strategy after comparing his utility against his random neighbour. This mechanism is usually called strategy selection and its details are shown in figure 1b. If the utility of his matched neighbour is higher than his utility, he will adopt his neighbour's strategy with a certain probability which is proportional to the utility difference between him and his matched neighbour. Beside strategy selection, we also include the mechanism of mutation where we flip the strategy of a randomly chosen individual after a certain period of time (see figure 1c). The mutation mechanism is necessary to avoid the system being trapped in the state of all agents adopting the same strategy.



**Figure 1. Model.** (a) Our model consists of both a social and an ecological system. The resource within the ecological system is increased through a linear resource inflow, and decreased via natural depreciation and human extraction. Each individual within the social system interacts based on a social network topology. Their payoffs depend on their ecological resource extraction. If the individual is a defector, an additional cost due to ostracism through social interaction with its neighbours has to be subtracted from his utility. A co-operator (blue) agrees to extract less resource according to the prior agreement and ostracizes any defectors that are connected to them. On the other hand, a defector (red) maximizes its payoff by extracting more than the agreement. The updating mechanism in (b) represents the process of selection. At each time step, a random individual compares his utility with that of a random neighbour. The probability of an individual changing his strategy to the opposite strategy is proportional to the utility difference. The updating mechanism in (c) represents the process of random mutation. At a certain mutation period, a random individual is selected to reverse its strategy (from co-operator to defector, or vice versa).

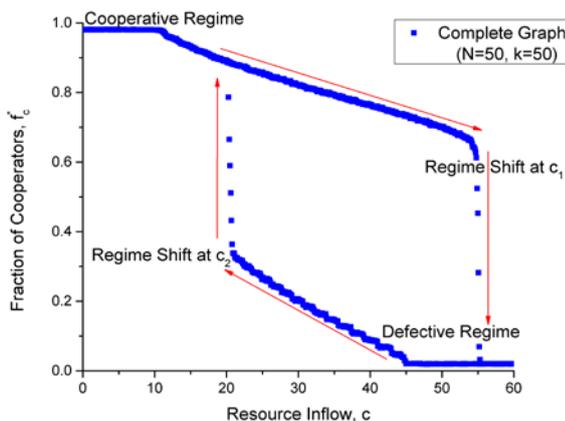
## 3. Methods

In this section we proceed to introduce the methods we employ to analyse our model. The analysis shall consist of two parts: the phenomenon of hysteresis with the presence of multiple stable states, and the early warning signals of the upcoming regime shifts. The coupling between the social and ecological aspects of the system creates a strong correlation between the fraction of co-

operators and the availability of resources, which lead to similar results for these two components. In consequence, we shall only focus on the social component of the system and drop the ecological part in our discussion.

### 3.1. Hysteresis and Multiple Stable States

Our system can fall into either a cooperative or a defective regime. In order to obtain a good approximation of the hysteresis cycle with multiple stable states, we average our results over many cycles. Note that a single cycle is defined as moving the socio-ecological states once around the hysteresis loop by adjusting the control parameter. For our studies, initial conditions are chosen such that the system begins at the cooperative regime. To ensure that the system is in the steady state, we evolve the system for a sufficiently long time before altering the value of our control parameter. For the first half of a hysteresis cycle, we increase the control parameter continuously and quasi-statically, driving the system gradually along the steady state values within a particular regime. The new state is then recorded when a new equilibrium is reached after each alteration. This process is repeated until a critical parameter is exceeded and the system undergoes a regime shift. We then reverse the process to evolve the system towards its initial state. Note that each complete cycle of parameter alteration gives a hysteresis curve.



**Figure 2. Hysteresis Structure.** The blue square symbol represents the average over 100 cycles of fraction of co-operators at equilibrium for different resource inflow  $c$ . The simulation starts at  $c=0$  and stops at  $c=60$ . The system is evolved at a fixed  $c$  for a sufficiently long time in order for the state to reach its equilibrium value. The parameter  $c$  is then increased by 0.1 and the process is repeated. After we have completed the computation of all equilibrium points between  $c=0$  and  $c=60$ , we reverse the process and determine the equilibrium points by varying  $c$  from 60 to 0. This whole process represents a single cycle of hysteresis. It is repeated and averaged over 100 cycles to produce the above average hysteresis cycle.

Figure 2 shows the hysteresis curve of a system based on social interactions from a complete graph of 50 individuals driven by a variation in the amount of resource inflow as our control parameter. Note that various control parameters whose adjustment can lead to the hysteresis cycle, but for simplicity, we shall only concern with

resource inflow as our control parameter in this paper. The result in Fig. 2 was averaged over 100 simulations. As the system becomes close to the first transition point  $c_1$ , a further increase in the amount of resource triggers a critical transition towards the defector equilibrium. Once the transition takes place, the previous states of the system cannot be restored through reversing the same path. During the second half of the hysteresis cycle, the cooperativeness of the population does not increase sharply back to its previous values at  $c_1$  as we decrease the amount of resource inflow. Instead, it increases slowly by tracing a distinct path before a second transition point  $c_2$  is reached. Then, a further reduction in the amount of resource inflow triggers another sharp transition: from the defective regime to the cooperative regime.

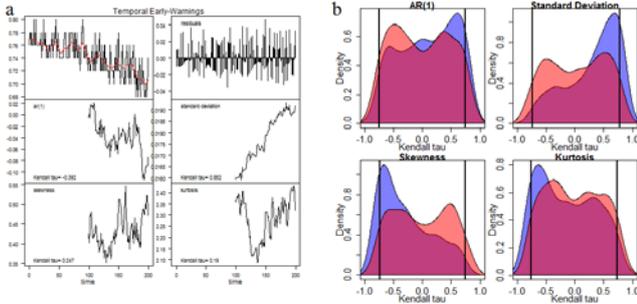
### 3.2. The Analysis of Early Warning Signals

Typically, early warning signals are obtained by exploiting the generic behaviour of the system close to critical transition, such as the phenomenon of critical slowing down. However, these early warning signals normally suffer from false detection. Previous research has provided a statistical comparison between the results from the test model and the null model to determine the accuracy of the model<sup>31,32,37</sup>. One of them is Receiver-Operating Characteristic (ROC) which is a very robust method for the evaluation of the performance of various indicators and to capture their trade-off between both false positive and false negative qualitatively<sup>32</sup>. However in this paper, in order to obtain the accuracy quantitatively, we compare the probability distribution of test model and null model directly by using p-value significant testing. The test model relates to the case where our control parameter increases very slowly till it reaches the tipping point (10 increments in 10,000 steps), i.e. we select the portion that precedes the potential transition. On the other hand, the null model is the situation without regime shift where the control parameter is kept fix, such that the system is only driven by stochastic fluctuations. The early warning signals can be obtained by means of either temporal patterns or spatial patterns. The accuracy of the early warning signals for different network degrees will be compared and discussed.

#### 3.2.1. Temporal Patterns

Temporal early warning signal is often handy because in most cases the time series data is the only information available to us. The data analysis that yields the early warning signals of interest usually requires several steps which include pre-processing, filtering, probing, and significance testing whose details can be found in Dakos et al<sup>21</sup>. In empirical observations, we are typically restrained by the frequency of observation (i.e. the time interval between points in the datasets). However, this does not happen in our case since our time series data arise from the model, and our results show that the accuracy of the early warning signals obtained is independent of the frequency of observation. Therefore, we only illustrate situations when the time

interval is 50. In this analysis, we use a rolling window with a size that is half of the whole time series datasets. We also filter the trends by using Gaussian smoothing to avoid spurious indications caused by the presence of strong local correlation structures in the time series (figure 3a, panel 1). The de-trended data (figure 3a, panel 2) is then analysed by several conventional indicators such as autocorrelation, standard deviation, skewness and kurtosis. We have quantified the indicator's trends by using the Kendall tau rank correlation which was computed through the R package: 'early warnings' (figure 3a, panel 3-6). To achieve the statistical comparison, we have replicated 1,000 realizations for each time series measurement from our simulation for both the case of test model and the null model. The distributions for both cases are compared to determine the accuracy of the early warning indicators (figure 3b, panel 1-4). Note that the vertical lines indicate the p value = 0.05 of the null model. Any value beyond these lines is considered significant. The accuracy of specific indicator is then quantified by the proportion of significant predictions attained against the total number of predictions attempted.

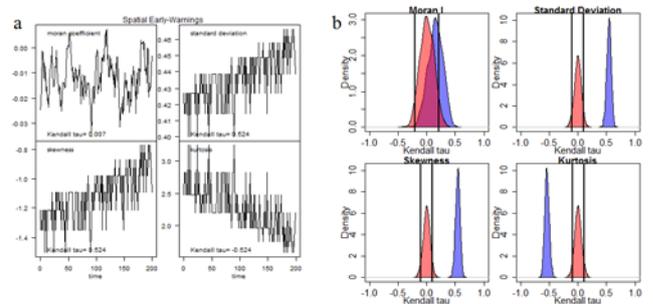


**Figure 3. Procedure for temporal early warning signals.** Figure (a) represents a single realization of early warning detection. Panel 1 gives the time series of the fraction of co-operators. The red curve is obtained after passing the time series through a Gaussian filter. Panel 2 shows the de-trended data that is employed for subsequent analysis. Each indicator is calculated within a rolling window that is half the size of the whole data. Panel 3-6 illustrates the trend for the following indicators: autocorrelation at lag-1, standard deviation, skewness, and kurtosis. The trend is then further analysed by means of the Kendall tau rank correlation. Figure (b) shows a comparison between the null model and the test model for each temporal indicator. The distribution in red represents the trend from the 1000 realizations of the null model. The vertical lines indicate the positions of the p-value = 0.05. On the other hand, the blue distribution represents the trend from the 1000 realizations of the test model.

### 3.2.2. Spatial Patterns

Spatial early warning signal is only useful if we have complete spatial information of the system. It often provides more accurate predictions in comparison to temporal early warning signal, although it is more difficult to exploit due to insufficient spatial data in many cases. In our work, the spatial pattern is obtained from the spatial distribution of each strategy (cooperative or defective) within the network structure. The spatial autocorrelation is quantified by means of the Moran spatial correlation,

$$= \frac{(N \sum_{i=1}^N \sum_{j=1}^N w_{ij} (x_i - \bar{x})(x_j - \bar{x}))}{(\sum_{i=1}^N \sum_{j=1}^N w_{ij}) \sum_{i=1}^N (x_i - \bar{x})^2}$$
, where  $w_{ij} = 1$  if node  $i$  and  $j$  are adjacent in their network structure, and  $w_{ij} = 0$  otherwise.  $i$  and  $j$  refer to the location of a node which represents an agent within the network structure. In our calculations, we let  $x_i = 1$  if agent  $i$  is using the cooperative strategy and  $x_i = 0$  if agent  $i$  is using the defective strategy. The standard deviation, skewness and kurtosis are modified into spatial measures as the second, third, and fourth moments about the spatial mean respectively<sup>38</sup>, i.e. the spatial variance is formally defined as  $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$ , the spatial skewness  $\gamma = \frac{1}{N} \sum_{i=1}^N \frac{(x_i - \bar{x})^3}{\sigma^3}$ , the second moment of spatial mean  $\kappa = \frac{1}{N} \sum_{i=1}^N \frac{(x_i - \bar{x})^4}{\sigma^4}$ . In many spatial early warning methods, a 2 dimensional space discretized into  $M$  and  $N$  units in  $x$  and  $y$  direction has been employed<sup>24,38</sup>. However, since the spatial structure in our paper is defined with respect to the agent's strategy which is organized in terms of network topology, we use the definition of local and global network statistics<sup>39</sup>. Similar to the temporal early warning method, we record the value of each indicators as the system approaches the critical transition. As the system gets closer to critical transition, we expect an increase in spatial autocorrelation and standard deviation. We shall quantify the trends exhibited through the spatial indicators by means of the Kendall tau rank correlation (figure 4a, panel 1-4). In order to achieve statistical comparison we have generated 1,000 realizations from time series simulation for the case of test model and also the null model. The test model is the trend of spatial indicator when the system is approaching critical transition and the null model is the trend of spatial indicator when the system is not approaching critical transition. These are illustrated in panel 1-4 of figure 4a. The distributions obtained for these two cases are then compared to determine its accuracy (figure 4b). Note that the vertical lines indicate the p value = 0.05 of the null model. Any value beyond these lines is considered as significant. Again, the accuracy of the specific indicator is quantified by the proportion of significant predictions achieved against the total number of predictions attempted.



**Figure 4. Procedure for spatial early warning signals.** Figure (a) illustrates a single realization of spatial early warning signals. Panel 1-4 give the trend of spatial autocorrelation (Moran coefficient), standard deviation, skewness, and kurtosis respectively. The trend is then analysed using the Kendall tau rank correlation. Figure (b) shows a comparison between the null model and the test model for different spatial indicators. The distribution

in red is obtained from the trend of 1000 realizations of the null model. The vertical lines show the position of the  $p$ -value = 0.05. The blue distribution is derived from the trend of 1000 realizations of the test model.

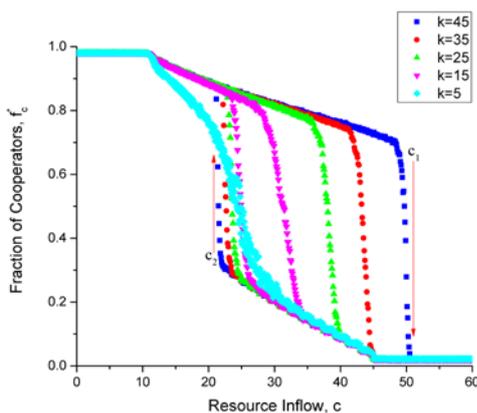
#### 4. Results

##### 4.1. The Effect of Network Properties on Hysteresis Structure.

In this section, we shall present our numerical results and discuss the effects of several network properties on the hysteresis structure of the system.

##### 4.1.1. Degree

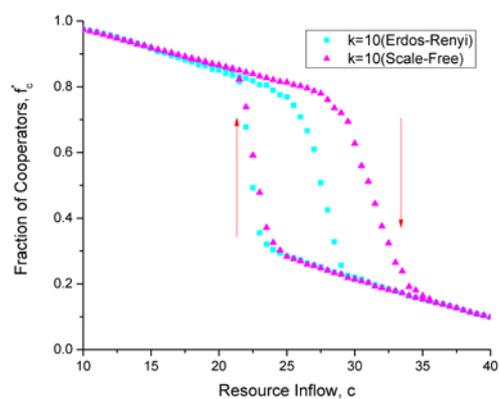
Some societies are more connected than others. Therefore, it is reasonable to study the effects of network degree and investigate its consequences on coupled socio-ecological systems. This has led us to employ the Erdos-Renyi graph with a size of  $N=50$  as our social network. We shall vary the average degree ( $k$ ) of this network to model societies with different average number of social connections. The resulting set of hysteresis structures obtained is presented in Fig. 5. We observe that as the average degree  $k$  decreases, the width ( $\Delta c = |c_1 - c_2|$ ) of the hysteresis curve reduces. As shown in Fig. 5, critical transitions happen around  $c_1=50$  and  $c_2=22$  for a population with  $k=45$ . When the network has a lower degree (for example  $k=25$ ), the regime shift towards the defective regime happens earlier (at  $c_1=40$ ) while the regime shift towards the cooperative regime occurs at a slightly larger value of resource inflow ( $c_2=25$ ). Interestingly, hysteresis effect is no longer observed for a population with very low number of social connection (i.e.  $k=5$ ). When the number of social ties is small, a reduction or increment of a single co-operator can have a large impact on the effectiveness of social ostracism within the local co-operator communities. In this case, the fraction of co-operators decreases faster as the control parameter increases and the system may regain its original state by following the same path as we reverse the process. On the other hand, when there are a large number of social connections, the reduction or increment of a single co-operator has relatively less impact on the effectiveness of social ostracism. Hence, there exists a critical point when social sanction can no longer hold the extra payoff offered by defective behaviour.



**Figure 5. The effects of network degree on hysteresis structure.** The simulation is based on the Erdos-Renyi random graph topology with network degrees ranging from  $k=45$  to  $k=5$  and a population size of  $N=50$ . We plot the fraction of co-operators versus resource inflow obtained from the simulation results for degree  $k=45$  (represented by blue square symbol) to  $k=5$  (represented by cyan diamond symbol) at a decrement of 5 unit each (see legend for the different colour and symbol). Note the reduction in hysteresis width as the degree is lowered.

##### 4.1.2. Topology

Most real world social networks are not random graphs. Here, we study the influence of network topologies on regime shifts in coupled socio-ecological system. Specifically, we compare the effects of two different network topologies: the Erdos-Renyi random network, and the scale-free network generated using the Chung-Lu algorithm<sup>40</sup> on the hysteresis structure. Note that we have raised the population size to  $N=200$  in order to enhance the effects from the scale-free network. Simulation results are shown in figure 6, where we noticed a difference in hysteresis width between the two hysteresis curves. Although these networks have the same degree, a scale-free network comprises a greater proportion of nodes with a larger degree. This feature boosts the effectiveness of the ostracism mechanism within the co-operative regime of a society with a scale-free network structures, such that the critical transition is prevented from happening earlier. Furthermore, since the effectiveness of the ostracism mechanism is dependent on the presence of a certain number of co-operators, the degree structure of the network has minimal effect when the state of the system is within the defecting regime. As a result, the critical transition from the defective regime into the cooperative regime occurs at similar control parameter value.

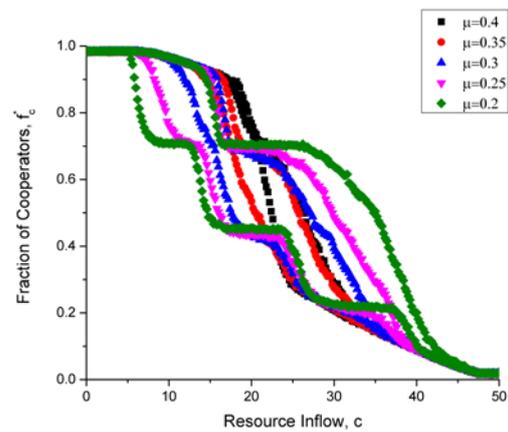
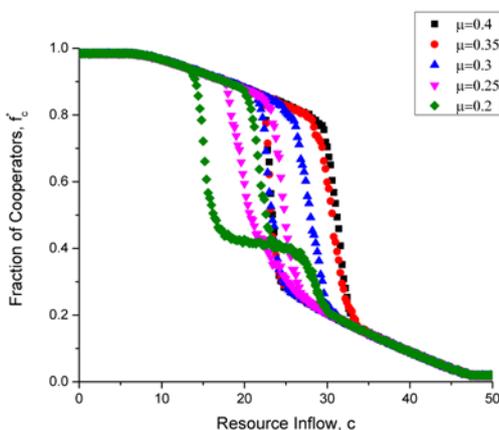


**Figure 6. The effects of network topology on hysteresis structure.** The cyan square represents the numerical results from the Erdos-Renyi random network topology, while the magenta triangle corresponds to that of a scale free network from the Chung-Lu algorithm. We have increased the population size to  $N=200$  in order to capture the scale free effect. Note that we have set the range of the control parameter  $c$  from  $c=10$  to  $c=40$ .

##### 4.1.3. Community Structure

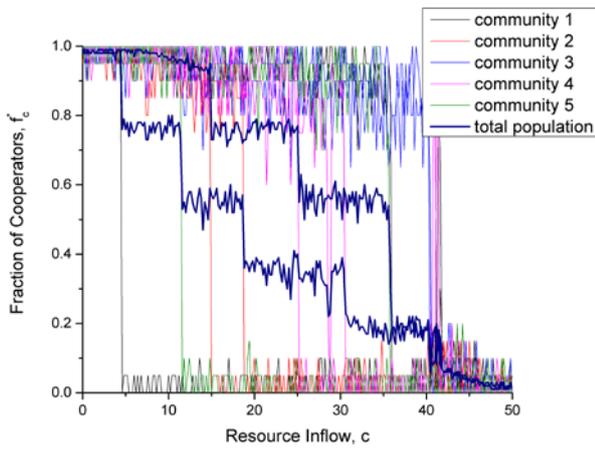
In many societies and global organization of the real world, social connection is found to consist of several structural subunits or communities associated with strongly interconnected components<sup>17</sup>. To study the effects of such social organizations with sub-structural units, we construct an artificial network with the properties of community structure. The community structure is created by rewiring the original Erdos-Renyi graph via a classification of all the nodes into several community groups. The quantifier “parameter mixing” ( $\mu$ ) measures the amount of intergroup connections with the total connections, and hence it indicates that each node shares a fraction  $\mu$  of its links with other community<sup>41</sup>. In this section, we consider a population size of  $N=100$ . This choice is motivated by practical considerations since 100 individuals can be easily divided into groups of 2, 4, and 5 with equal number of individuals within each community group.

Our simulation results are shown in figure 7, where we observe a change in the hysteresis structures as  $\mu$  varies. These figures illustrate the case of a network with degree  $k=15$  consisting of 2 and 4 community groups. We observe that as  $\mu$  is lowered (modularity increased), the regime shift tends to occur earlier from the cooperative regime to the defective regime. This effect is observed for any number of community groups (here we display the results for 2 and 4 community groups). In the case of low  $\mu$ , the ostracism mechanism operates mainly within the individual community which tends to isolate from each other due to the stronger intra-group connections. Since ostracism functions via the presence of co-operators, its effectiveness reduces as the co-operators become more isolated within each separate group. In consequence, the regime shift to the defective regime occurs earlier. Furthermore, for very low  $\mu$ , we observe the occurrence of multiple hysteresis. Instead of a total collapse or a total revival, the system is found to collapse or revive step by step. From the plot, we observe that the multiple hysteresis and the steps of the regime shifts are obscured in lieu of the averaging effect. Several tipping points of the last few steps have been averaged out and seem to have become mixed into a single shift.



**Figure 7. The effects of community structure on the hysteresis curve.** The simulation is based on a modified random graph with a topology that contains community structure. Note that the network is fixed with a degree of  $k=15$  and a population size of  $N=100$ . The results are obtained from simulations based on mixing parameters that range from  $\mu = 0.4$  (represented by square symbol) to  $\mu = 0.2$  (represented by diamond symbol) with a decrement of 0.05 unit (see legend for the different symbols) for (a) 2 community groups; and (b) 4 community groups.

In order to gain a better picture on the reasons behind the multiple hysteresis phenomena, we have plotted single realizations of hysteresis cycle in a network structure with 5 community groups for  $\mu = 0.2$ . We have plotted the fraction of co-operators within each community as well as that within the whole society. The results show that as the control parameter increases, the cooperative behaviour does not collapse globally but instead locally within the community. More precisely, we can see from figure 8 that the cooperativeness within community 2 collapses first while those of other communities continue to survive. As the control parameter is further increased, community 4 is observed to collapse next. This is followed by community 1 and then community 5. Finally, community 3 collapses. Interestingly, the network community structure prevents the ostracism mechanism to act effectively across communities and thus prevent the concomitant collapse of cooperative behaviour across the whole society. It is interesting that the reversal of the communities from the defector regime to the cooperation regime does not necessarily follow the same sequence as that when the cooperation of the communities collapses. By combining all the hysteresis structures of each community, we then observe the multiple hysteresis of the whole population as represented in bold colour in figure 8.



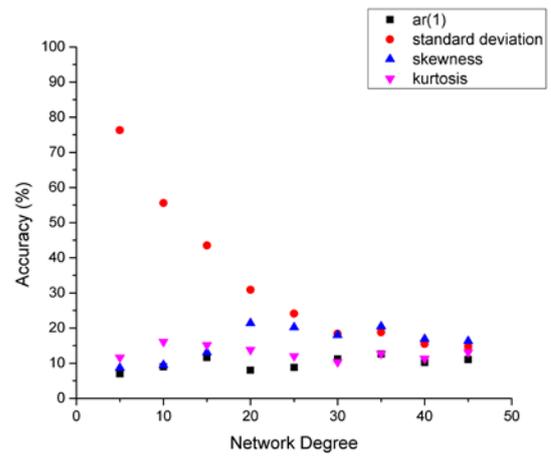
**Figure 8. Single realization of the effects of community structure on the hysteresis curve.** The simulation is performed according to a modified random graph with the topology of community structure. Note that the network is fixed with a degree of  $k=15$ , a population size of  $N=100$ , and a mixing parameter of  $\mu=0.2$ . We have plotted the fraction of co-operators against resource inflow for each community which is represented by curves of different colour. The colour in bold is for the case when we consider the whole population.

#### 4.2. Early Warning Signals.

In this section, we shall show our numerical results and discuss the effects of network connectivity on the accuracy of several conventional early warning signals. Note that all the necessary details with regards to the early warning signals have already been discussed in the methods section.

##### 4.2.1. Temporal Patterns

Figure 9 illustrates the accuracy of specific temporal indicators as the network degree varies. Our results here show the percentage of significant correct prediction. We found that most of the temporal early warning signals are not accurate enough to predict the approaching transition. From the figure, we can see that only temporal standard deviation is sensitive enough to detect future regime shift. Other indicators are not able to distinguish between the null model (stable system) and the test model (system approaching critical transition). We observe that the accuracy of the temporal standard deviation increases as the network degree is lowered. This results from the following. For the case of high degree network, the stability of the regime reduces gradually as the control parameter increases. On the other hand, the stability drops faster in the case of low degree network as the control parameter increases. Since the dynamics of the system fluctuates more rapidly (higher variance) when the system is unstable, the temporal standard deviation is able to capture the increasing trend of variance more effectively. This explains why the temporal standard deviation performs better when the network degree is low versus that when the network degree is high.



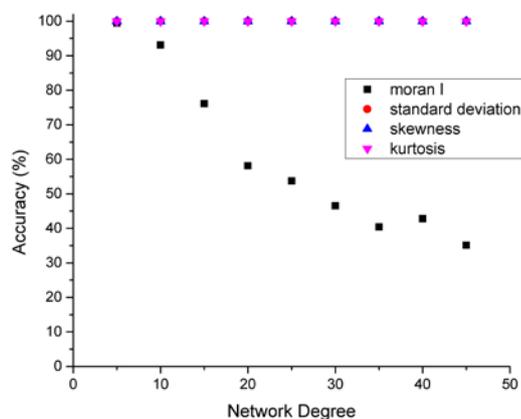
**Figure 9. Accuracy of temporal early warning signals.** A plot on the accuracy of different temporal early warning indicators: autocorrelation at lag-1, standard deviation, skewness, and kurtosis (see legend), versus the network degree. Note that each point is obtained after determining the percentage of significant accurate predictions through a comparison made against the null model.

##### 4.2.2. Spatial Patterns

Figure 10 shows the accuracy of specific spatial indicators (Moran I, standard deviation, skewness, kurtosis) as the network degree changes. In comparison to the temporal early warning signals, we found that most of the spatial indicators are sensitive enough to predict the occurrence of an approaching regime shift. In fact, the second, third, and fourth moment of the spatial mean (i.e. the standard deviation, skewness, and kurtosis) are able to predict the approaching regime shift with 100% accuracy. For these indicators, the distribution of the null model and the test model is found to be totally separated without any overlap (see figure 4b). The separation between the test model and the null model indicates the underlying accuracy of the model since the early warning signal can be clearly distinguished from the false signal. Far away from the critical transition, only the strategy of mutation is dominant in the updating mechanism. Therefore, the strategy of each node varies near the spatial mean of the null model, as indicated by the second, third, and fourth spatial moments. On the other hand, since the strategy of selection is dominant in the test model, the strategy of each node is observed to vary far from the spatial mean.

In the case of spatial correlation, the accuracy of the Moran I is found to increase as the network degree is lowered, and it can predict with 100% accuracy when the degree is very low. This can be understood as follow. Near the critical transition, we can perceive that every part of the network becomes spatially more similar to each other. For high degree network, since everyone is almost connected to everyone else, the system is already spatially correlated even for the null model. This makes it difficult to distinguish between the null and the test model. Such spatial correlation reduces as the network degree decreases, thus enabling the null model to be distinguishable from the test model. Spatial early warning signals are found to be more robust for the

detection of approaching regime shift compared to temporal early warning signals. In reality, however, this approach is very difficult to achieve practically because there is a need to have complete spatial strategy information of every individual. Perhaps in the future, it is possible to obtain the spatial behaviour and social network of every individual through big data by means of smart phone location tracking or social networking services.



**Figure 10. Accuracy of spatial early warning signals.** A plot on the accuracy of various spatial early warning indicators: Moran coefficient, spatial standard deviation, spatial skewness, and spatial kurtosis (see legend), versus the network degree. Note that each point is obtained after determining the percentage of significant accurate predictions through a comparison made against the null model.

## 5. Conclusion

We have investigated into the influence of complex social network interactions on regime shifts in coupled socio-ecological system as well as our ability to make accurate prediction on its occurrence. We have based our study using the TSL model, with the inclusion of social interactions modelled by different network topologies and employing a discrete choice mechanism to update agent's strategy which involves selection (utility driven strategy selection) and mutation (random strategy updating). Our results show that intrinsic social network properties can yield interesting multi-stable hysteresis structures, and can also have subtle effects on the accuracy of early warning signals. Thus, a more detailed understanding on the social interaction network properties as well as the associated socio-ecological parameters within a society would provide deeper insights that will be important for its proper protection. More importantly, it will enable us to improve our abilities to anticipate or even avoid the unsought for regime shift that can be catastrophic. In consequence, we perceive that the results of this work would be especially relevant and beneficial for decision making and management planning within the field of coupled socio-ecological systems.

## References

1. Hoekstra, A. Y. Water scarcity challenges to business. *Nat. Clim. Change* **4**, 318–320 (2014).

2. Problems of Scarcity and Pollution. *Nature* **230**, 543–543 (1971).
3. Jarvis, A. J., Leedal, D. T. & Hewitt, C. N. Climate-society feedbacks and the avoidance of dangerous climate change. *Nat. Clim. Change* **2**, 668–671 (2012).
4. Parmesan, C. & Yohe, G. A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**, 37–42 (2003).
5. ROSENBERG, D. M., MCCULLY, P. & PRINGLE, C. M. Global-Scale Environmental Effects of Hydrological Alterations: Introduction. *BioScience* **50**, 746–751 (2000).
6. West German forests: Deterioration, but some recovery. *Nature* **319**, 529–529 (1986).
7. Miller, G. H. *et al.* Ecosystem collapse in Pleistocene Australia and a human role in megafaunal extinction. *Science* **309**, 287–290 (2005).
8. Steffen, W., Crutzen, P. J. & McNeill, J. R. The Anthropocene: Are Humans Now Overwhelming the Great Forces of Nature. *AMBIO J. Hum. Environ.* **36**, 614–621 (2007).
9. Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413**, 591–596 (2001).
10. Dent, C. L., Cumming, G. S. & Carpenter, S. R. Multiple states in river and lake ecosystems. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **357**, 635–645 (2002).
11. Scheffer, M. & Carpenter, S. R. Catastrophic regime shifts in ecosystems: linking theory to observation. *Trends Ecol. Evol.* **18**, 648–656 (2003).
12. Scheffer, M. *et al.* Anticipating Critical Transitions. *Science* **338**, 344–348 (2012).
13. Nowak, M. A. & May, R. M. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992).
14. Lieberman, E., Hauert, C. & Nowak, M. A. Evolutionary dynamics on graphs. *Nature* **433**, 312–316 (2005).
15. Ohtsuki, H., Hauert, C., Lieberman, E. & Nowak, M. A. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* **441**, 502–505 (2006).
16. Nowak, M. A., Tarnita, C. E. & Antal, T. Evolutionary dynamics in structured populations. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 19–30 (2010).
17. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
18. Weng, L., Menczer, F. & Ahn, Y.-Y. Virality Prediction and Community Structure in Social Networks. *Sci. Rep.* **3**, (2013).
19. Scheffer, M. Complex systems: Foreseeing tipping points. *Nature* **467**, 411–412 (2010).
20. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
21. Dakos, V. *et al.* Methods for Detecting Early Warnings of Critical Transitions in Time Series Illustrated Using Simulated Ecological Data. *PLoS ONE* **7**, e41010 (2012).
22. Dakos, V., van Nes, E. H., D'Odorico, P. & Scheffer, M. Robustness of variance and autocorrelation as indicators of critical slowing down. *Ecology* **93**, 264–271 (2012).
23. Dakos, V., Nes, E. H. van & Scheffer, M. Flickering as an early warning signal. *Theor. Ecol.* **6**, 309–317 (2013).
24. Dakos, V., Nes, E. H. van, Donangelo, R., Fort, H. & Scheffer, M. Spatial correlation as leading indicator of catastrophic shifts. *Theor. Ecol.* **3**, 163–174 (2009).
25. Livina, V. N. & Lenton, T. M. A modified method for detecting incipient bifurcations in a dynamical system. *Geophys. Res. Lett.* **34**, L03712 (2007).
26. Carpenter, S. R. & Brock, W. A. Rising variance: a leading indicator of ecological transition. *Ecol. Lett.* **9**, 311–318 (2006).
27. Seekell, D. A., Carpenter, Stephen R. & Pace, M. L. Conditional Heteroscedasticity as a Leading Indicator of Ecological Regime Shifts. *Am. Nat.* **178**, 442–451 (2011).
28. Held, H. & Kleinen, T. Detection of climate system bifurcations by degenerate fingerprinting. *Geophys. Res. Lett.* **31**, L23207 (2004).
29. Biggs, R., Carpenter, S. R. & Brock, W. A. Turning back from the brink: detecting an impending regime shift in time to avert it. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 826–831 (2009).

30. Hastings, A. & Wysham, D. B. Regime shifts in ecological systems can occur with no warning. *Ecol. Lett.* **13**, 464–472 (2010).
31. Boettiger, C. & Hastings, A. Early Warning Signals and the Prosecutor’s Fallacy. *Proc. Roy. Soc. B.* **279**, 4734–4739 (2012).
32. Boettiger, C. & Hastings, A. Quantifying Limits to Detection of Early Warning for Critical Transitions. *J. Roy. Soc. Interface.* **9**, 2527–2539 (2012).
33. Hardin, G. The Tragedy of the Commons. *Science* **162**, 1243–1248 (1968).
34. Sethi, R. & Somanathan, E. The Evolution of Social Norms in Common Property Resource Use. *Am. Econ. Rev.* **86**, 766–88 (1996).
35. Noailly, J., Withagen, C. A. & Bergh, J. C. J. M. van den. Spatial Evolution of Social Norms in a Common-Pool Resource Game. *Environ. Resour. Econ.* **36**, 113–141 (2007).
36. Tavoni, A., Schlüter, M. & Levin, S. The survival of the conformist: social pressure and renewable resource management. *J. Theor. Biol.* **299**, 152–161 (2012).
37. Carl Boettiger, A. H. No early warning signals for stochastic transitions: insights from large deviation theory. *Proc. Biol. Sci.* **280**, 20131372 (2013).
38. Kéfi, S. *et al.* Early Warning Signals of Ecological Transitions: Methods for Spatial Patterns. *PLoS ONE* **9**, e92097 (2014).
39. Okabe, A. & Sugihara, K. in *Spatial Analysis along Networks* 137–151 (John Wiley & Sons, Ltd, 2012).
40. Chung, F. & Lu, L. Connected Components in Random Graphs with Given Expected Degree Sequences. *Ann. Comb.* **6**, 125–145 (2002).
41. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**, 046110 (2008).

# Progress in Ultrasonic Nano Manipulations

Junhui Hu\*, Qiang Tang, Xu Wang, Xiaofei Wang

**Abstract**—Ultrasonic nano manipulation is an emerging technology, which has great potential applications in the assembly, measurement and fabrication of nano materials, handling of biological samples, manufacturing of nano sensors, new material syntheses, etc. In recent three years, the author's research team proposed and developed a series ultrasonic manipulators with the functions such as nano trapping and transfer, nano rotary driving, and nano concentration. Controlled acoustic streaming eddies are used in the nano manipulations. Compared with other nano manipulation techniques, they have the features such as very low temperature rise at the manipulation area, little selectivity to manipulated samples, being implemented on the substrates given by customers. etc. This paper reports our latest progress in the function enhancement of ultrasonic nano manipulations, simulation of the acoustic streaming employed, and modeling of the ultrasonic devices.

**Keywords**—Nano manipulation, Acoustic streaming, Ultrasonic device.

## I. INTRODUCTION

With the development of biomedicine, micro/nano fabrication, new material and so on, devices for actuating nano materials are being required [1, 2]. Required actuation functions for nano materials include trapping, positioning, transfer, release, revolution, removal, concentration, assembly, sorting, etc. These functions are also called nano manipulation. However, most of the above listed nano manipulation functions cannot be effectively and efficiently realized by the conventional actuation technology, which have limited driving forms and operating principles [2]. To fulfill the demands, lots of strategies have been proposed and investigated. They can be classified as optical [1, 3], magnetic [4], electric [5], mechanical [6], AFM [7], microfluidic [8] and acoustic methods [9-15], based on the physical principles which they use.

Ultrasonic nano manipulations utilize the sound induced

This work is supported by the following funding organizations in China: National Science Foundation of China (No. 91123020), State Key lab of Mechanics and Control of Mechanical Structures (MCMS-0313G01 and MCMS-0314G01), Nanjing University of Aeronautics and Astronautics (No. S0896-013), the "111" project (No. B12021), and PAPD.

Junhui Hu is with State Key Lab of Mechanics and Control of Mechanical Structures, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (corresponding author to provide phone: 18912946712; e-mail: ejhhu@nuaa.edu.cn).

Qiang Tang is with State Key Lab of Mechanics and Control of Mechanical Structures, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: tangqiang102@126.com).

Xu Wang is with State Key Lab of Mechanics and Control of Mechanical Structures, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: zixu630296301@126.com).

Xiaofei Wang is with State Key Lab of Mechanics and Control of Mechanical Structures, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: 1054032135@qq.com).

flow or acoustic streaming to manipulate nanoscale materials. In recent three years, the authors' group proposed and realized a series of nano manipulations by the means of controlled acoustic streaming. They include trapping, orientation, positioning, transfer and rotation of individual nanowires in deionized water, and concentration of nanowires and nanoparticles in deionized water [2, 9-13]. It has the features such as little selectivity to the material properties of manipulated samples, little heat damage to manipulated samples, diverse manipulation functions, and no need to dispose MEMS or NEMS structures on the substrate. Although it has very large potential applications in the fields such as biomedicine, micro/nano fabrication, material engineering, renewable energy, etc., researches on the principle, structure design, and application of these devices are still superficial and insufficient [2]. Actually there were few reports on the ultrasonic manipulations of a single nano object before the authors' work. This paper reports our latest progress in the function enhancement of ultrasonic nano manipulations, simulation of the acoustic streaming employed, and modeling of the ultrasonic devices.

## II. INTEGRATION OF NONCONTACT AND CONTACT TRAPPING FUNCTIONS INTO ONE DEVICE

In ultrasonic nano trapping, there are two working modes, i.e., the noncontact and contact modes. The noncontact trapping mode enables the device to handle sticky nano samples, and the contact trapping mode makes the transfer of a trapped sample convenient. However, the existing technology cannot integrate the noncontact and contact nano trapping functions into one device [9, 12].

Fig. 1 shows the experimental setup to implement the noncontact and contact-type trapping of individual nanowires by one device. The device is simply made up of the piezoelectric plate, vibration transmission needle (VTN) made of steel, and micro manipulating probe (MMP) made of fiberglass. The VTN is bonded along the narrow side of the piezoelectric plate. The MMP is bonded to the VTN's tip, and parallel to the piezoelectric plate. The resonance frequency of the device is about 136 kHz, at which the VTN vibrates flexurally. In the frequency range from 131.2 ~ 132.2 kHz, the trapped nanowire is not in contact with the MMP, and it is in contact with the MMP in the frequency range from 133.9 ~ 134 kHz. Figs. 2 and 3 contain a series of images to show the noncontact and contact trapping and transfer of a silver nanowire, respectively. In both modes, the AgNW rotates while being sucked to the MMP. From image *b* to *d* in Fig. 2, the trapped nanowire is moved on the substrate surface by moving the manipulating device. From image *d* to *g* in Fig. 3, the trapped wire is moved above the substrate surface, and in image *h* in Fig. 3, the trapped nano wire is released.

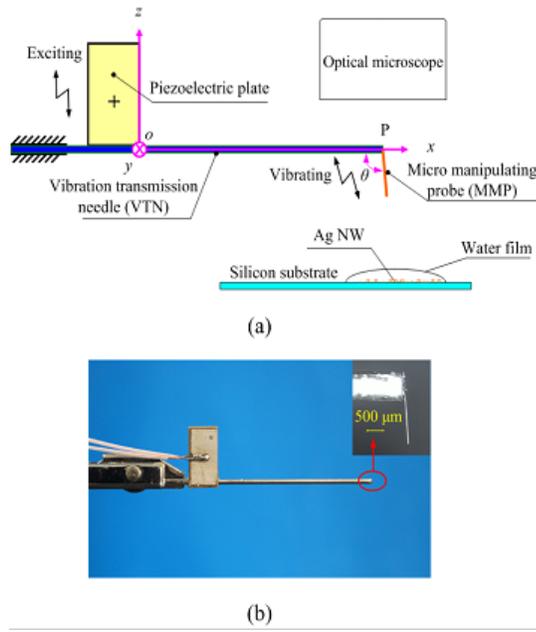


Fig. 1 Experimental setup for the noncontact and contact-type trapping of a single silver nanowire. (a) Schematic diagram. (b) Construction of the ultrasonic transducer.

The noncontact and contact trapping modes are realized by employing different acoustic streaming field patterns around the micro manipulating probe. Our calculation shows that the difference in acoustic streaming fields in the noncontact and contact modes, is caused by the change of the phase difference among the normal vibration components at the root of the micro manipulating probe.

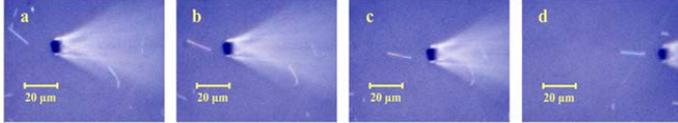


Fig. 2 Noncontact trapping of a single AgNW by the MMP's tip in water film.

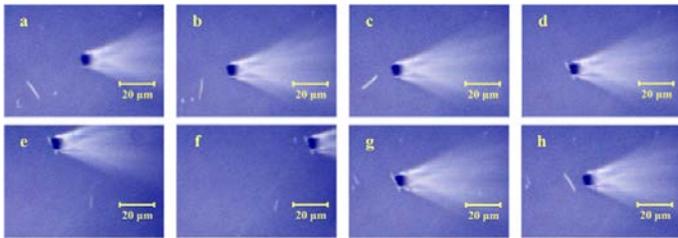


Fig. 3 Contact trapping of a single AgNW by the MMP's tip in water film.

Our experiments show that the noncontact mode has a working frequency band width of 1 kHz, while the contact mode has a working frequency band width of only 0.1 kHz. Increasing the working frequency band width for the contact mode remains a challenge.

### III. ACOUSTIC STREAMING

At the present stage, acoustic streaming is the only means employed in the ultrasonic nano manipulations [2]. For better and wider applications of acoustic streaming in nano manipulation, more convenient and efficient numerical methods are needed to calculate the acoustic streaming field in the devices and to analyze its change with the working and structural parameters of devices [16]. We proposed and developed a numerical method, which can make use of the COMSOL Multiphysics finite element method (FEM) software to effectively simulate the acoustic streaming. Furthermore, based on the simulation results, effective methods for controlling the acoustic streaming fields in nano manipulations have been achieved.

The computation process consists of three steps [16]. In the first step, the sound field is solved with the multiphysics coupling modules of the software. In the second step, vibration velocity and sound pressure of the sound field are used to calculate spatial gradients of the Reynolds stress and mean pressure, which generate the acoustic streaming, by the post processing functions of the software. In the last step, the steady acoustic streaming is solved by the fluidic dynamics module, with proper boundary conditions for the acoustic streaming. The steady acoustic streaming satisfies the following equation:

$$\rho_0(\bar{u}_i \partial \bar{u}_j / \partial x_i) = F_j - \partial \bar{p}_2 / \partial x_j + \eta \nabla^2 \bar{u}_j \quad (1)$$

where  $\bar{u}_i$  is acoustic streaming velocity, repeated suffix  $i$  and  $j$  represent  $x$ ,  $y$  and  $z$  in a 3D model,  $\rho_0$  is the medium density in the undisturbed state,  $F_j$  is the gradient of the Reynolds stress which acts on the fluid as a driving force of the acoustic streaming, and  $\bar{p}_2$  is the time average of the 2nd order pressure or mean pressure.  $F_j$  is calculated by

$$F_j = -\partial(\overline{\rho_0 u_i u_j}) / \partial x_i \quad (2)$$

where  $u_i$  is the vibration velocities in the sound wave, and the bar signifies the mean value over one period.  $\bar{p}_2$  is calculated by

$$\bar{p}_2 = \frac{1}{2\rho_0 c_0^2} \frac{B}{A} \langle p_1^2 \rangle \quad (3)$$

where  $p_1$  represents the (1st order) sound pressure,  $\langle \rangle$  represents the time average over one time period,  $c_0$  is the medium sound speed in the undisturbed state, and  $\frac{B}{A}$  is the nonlinear parameter of the medium. The acoustic streaming also satisfies the continuity equation

$$\rho_0 \partial \bar{u}_i / \partial x_i = 0 \quad (4)$$

Fig. 4(a) shows the contact type trapping process of a AgNW on the surface of a silicon substrate in deionized water film, reported in Ref. 9, and Fig. 4(b) is the calculated acoustic streaming field on the silicon substrate surface and in the  $yz$  vibration plane. At the root of the micro manipulation probe (the excited part), there are three orthogonal vibration components, which have different amplitudes and initial phases. According to our calculation, the acoustic streaming pattern is dependent on the phase differences and vibration amplitudes of these three orthogonal components. To generate

a useful acoustic streaming field for the contact type trapping of a nanowire, the phase difference between the  $y$  (or  $x$ ) and  $z$  vibration components must be close to  $\pm 90^\circ$ , and the amplitude of the  $x$  (or  $y$ ) vibration component must be small enough compared to the other vibration components.

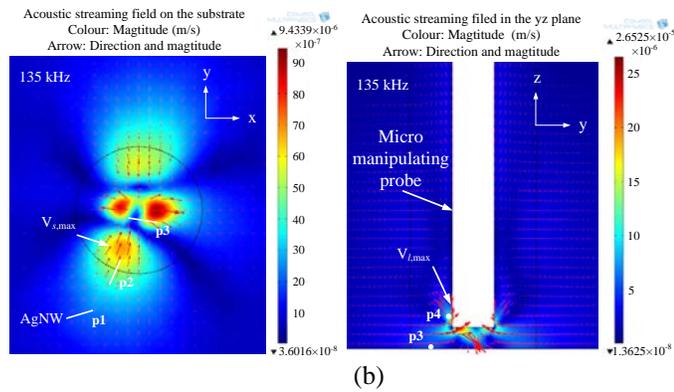
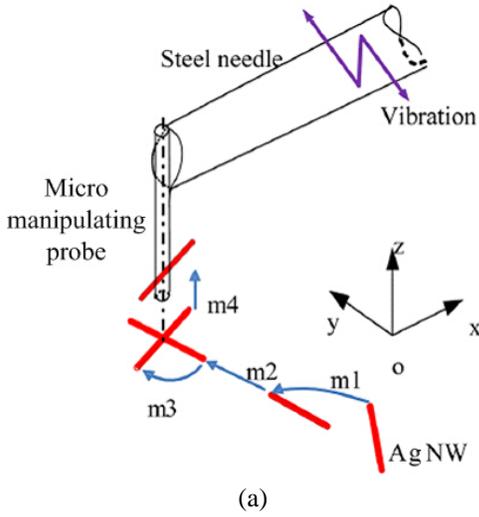


Fig. 4. Simulated acoustic streaming field employed by the contact type nanowire trapping. (a) A schematic diagram of the driving and trapping process for a AgNW, in which the micro manipulating probe is in the  $yz$  plane which is perpendicular to the steel needle. (b) Simulated acoustic streaming fields on the substrate surface (left) and in the  $yz$  plane (right).

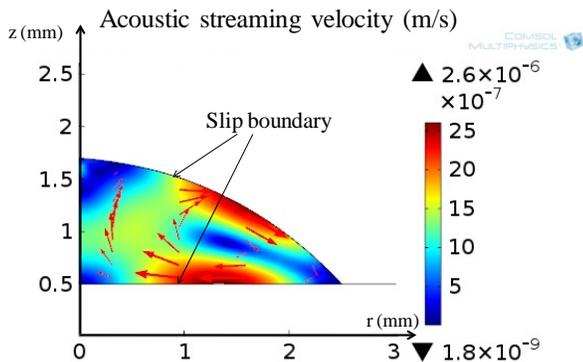


Fig. 5 Computed acoustic streaming field in the droplet on an ultrasonic stage.

Also, a droplet-ultrasonic stage system, in which a micro droplet located at the center of the ultrasonic stage is used to concentrate nanoscale material [10], is modeled and analyzed by the FEM, as shown in Fig. 5. The computed acoustic streaming field, shown in Fig. 5, can well explain the nano concentration phenomenon in the droplet-ultrasonic stage system, and useful guidelines for enhancing the concentration capability without sacrificing the manipulation stability are also obtained.

#### IV. DEVICE MODELING

Controlled rotary driving of single nano objects is an important technology in the assembling of nano structures, handling of biological samples, nano measurement, etc [2]. However, there have been little analyses on the ultrasonic transducers for the ultrasonic nano rotary driving [13], which makes the transducer's optimization impossible. Recently, the vibration characteristics of the ultrasonic transducer for rotary driving of single nanowires (NWs), which has been proposed by the authors' group, have been analyzed by the 3D finite element method (FEM), and some useful guidelines for designing the transducer are achieved.

Fig. 6 shows the structure and size of the vibration excitation system. The ANSYS software is used in the FEM analyses. A 3D FEM model of the device is shown in Fig. 7. The solid5 elements are used for the ceramics and the solid45 elements elsewhere; A constant damping ratio and the Full Method solver are used for the harmonic response calculation.

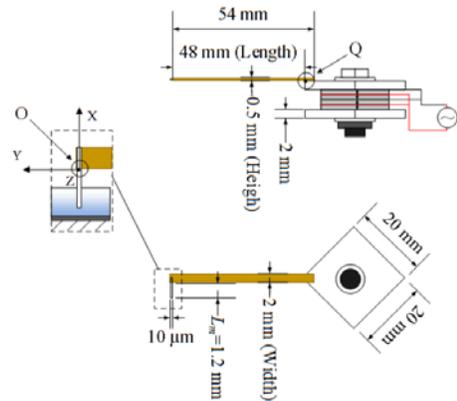


Fig. 6 Experimental setup and the ultrasonic device for the rotary driving of a single AgNW in water film on a silicon substrate.

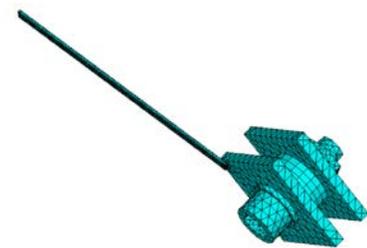


Fig. 7 A 3D FEM mesh model of the device.

The phase of the  $Y$ -directional vibration displacement minus that of the  $Z$ -directional vibration displacement at point  $O$  is defined as  $\Delta\phi_O$ . Fig. 8 shows the computed  $\Delta\phi_O$  versus driving frequency. It is seen that there exist some frequencies at which  $\Delta\phi_O = \pm 90^\circ$ , which means that the resultant of the  $Y$ - and  $Z$ -directional vibration components of the micro manipulating probe (MMP) is an elliptical motion at these frequencies. Thus at these driving frequencies, eddies can be generated around the MMP, which can drive the NWs to rotate. This well explains the experimental phenomenon reported in our previous work [13]. Moreover, based on the order of magnitude, it is known that point  $A$  corresponds to the working point in the experiments. Fig. 9 shows the computed vibration displacement at the MMP's tip versus the MMP's length  $L_m$ . It is seen that at 137 kHz, the MMP with a length  $L_m$  of 1.42 mm resonates. To ensure the performance consistency of the device, the MMP's length  $L_m$  or the driving frequency should be designed to avoid the resonance of the MMP. In addition, it is found that the working point can still exist when the commonly used metal materials in ultrasonic transducers, such as steel, copper and aluminum, are used as the vibration transmission strip, and may become unstable or disappears when the vibration transmission strip's length, width and height changes.

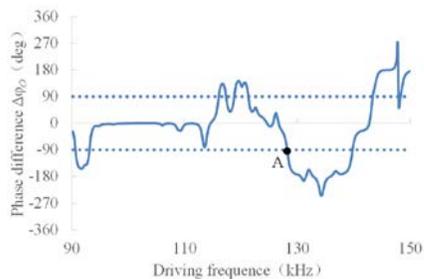


Fig. 8 Computed phase difference  $\Delta\phi_O$  versus driving frequency.

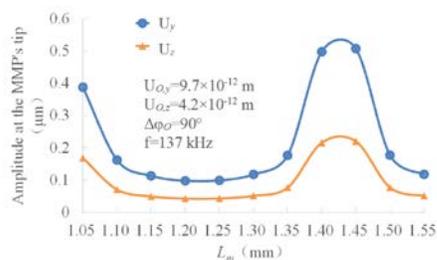


Fig. 9 Computed vibration displacement at the MMP's tip versus the MMP's length  $L_m$ .

## V. SUMMARY

The experimental and theoretical work has demonstrated that acoustic streaming can be used as an effective physical means for nano manipulations, and it can be effectively controlled by the phase difference between the normal vibration components of the micro manipulating probe. The noncontact and contact trapping functions can be integrated in one device by utilizing two different acoustic streaming fields generated at different working frequencies. Vibration control of the ultrasonic

manipulators is critical to realize or enhance a nano manipulation function. As an emerging actuating technology, the ultrasonic nano manipulation is facing lots of technological challenges such as the diversification of manipulation functions and manipulated samples, enhancement of manipulation functions, device vibration control, etc.

## REFERENCES

- [1] A. Ashkin, *Optical Trapping & Manipulation of Neutral Particles Using Lasers*. Singapore: World Scientific Publishing, Dec. 2006.
- [2] J. Hu, *Ultrasonic Micro/Nano Manipulations*, Singapore: World Scientific Publishing, April 2014.
- [3] A. Ashkin, Acceleration and trapping of particles by radiation pressure, *Phys. Rev. Lett.*, vol. 24, pp. 156–159, 1970.
- [4] M. Tanase, L. A. Bauer, A. Hultgren, D. M. Silevitch, L. Sun, D. H. Reich, P. C. Searson, and G. J. Meyer, Magnetic alignment of fluorescent nanowire, *Nano Lett.*, vol.1 (3), pp. 155–158, 2001.
- [5] J. Castillo, M. Dimaki, and W. E. Svendsen, Manipulation of biological samples using micro and nano techniques, *Integr. Biol.*, 1, pp. 30–42, 2009.
- [6] K. Molhave, T. Wich, A. Kortschack, and P. Boggild, Pick-and-place nanomanipulation using microfabricated grippers, *Nature Nanotech.*, vol. 17 (10), pp. 2434, 2006.
- [7] M. Sitti, B. Aruk, K. Shintani, and H. Hashimoto, Scaled teleoperation system for nano-scale interaction and manipulation, *Advanced Robotics*, vol. 17 (3), pp. 275–291, 2003.
- [8] M. Lu, S. Yang, Y. Ho, C. L. Grigsby, K. W. Leong, and T. Huang, Shape-Controlled Synthesis of Hybrid Nanomaterials via Three-Dimensional Hydrodynamic Focusing, *ACS Nano*, 10.1021/nn502549v.
- [9] N. Li, J. Hu, H. Li, S. Bhuyan, and Y. Zhou, Mobile acoustic streaming based trapping and 3-dimensional transfer of a single nanowire, *Appl. Phys. Lett.*, vol.101 (9), pp. 093113, 2010.
- [10] Y. Zhou, J. Hu, S. Bhuyan, "Manipulations of Silver Nanowires in a Droplet on Low-Frequency Ultrasonic Stage", *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol.60, no. 3, pp.622-629, 2013.
- [11] B. Yang and J. Hu, Linear Concentration of Microscale Samples under an Ultrasonically Vibrating Needle in Water on a Substrate Surface, *Sensors and Actuators B*, vol. 193, pp. 472–477, 2014.
- [12] H. Li, J. Hu, Noncontact Manipulations of a Single Nanowire Using an Ultrasonic Micro-Beak, *IEEE Transactions on Nanotechnology*, vol.13, no. 3, pp.469-474, May 2014.
- [13] N. Li, J. Hu, Sound Controlled Rotary Driving of a Single Nanowire, *IEEE Transactions on Nanotechnology*, vol. 13, no. 3, pp. 437-441, May 2014.
- [14] A. L. Balk, L. O. Mair, P. P. Mathai, P. N. Patrone, W. Wang, S. Ahmed, T. E. Mallouk, J. A. Liddle, and S. M. Stavis, Kilohertz Rotation of Nanorods Propelled by Ultrasound, Traced by Microvortex Advection of Nanoparticles. *ACS Nano* 8, pp. 8300-8309, 2014.
- [15] S. Ahmed, D. T. Gentekos, C. A. Fink, and T. E. Mallouk, Self-Assembly of Nanorod Motors into Geometrically Regular Multimers and Their Propulsion by Ultrasound, *ACS Nano* 10.1021/nn5039614.
- [16] Q. Tang, J. Hu, Diversity of Acoustic Streaming in a Rectangular Acoustofluidic Field, *Ultrasonics*, vol. 58, pp. 27-34, 2015.

**Junhui Hu** received his Ph.D. Degree from Tokyo Institute of Technology, Tokyo, Japan, in 1997, and B. E. and M. E. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 1986 and 1989, respectively. He is a Chang-Jiang Distinguished Professor of the Ministry of Education, China, the director of Precision Driving Lab at Nanjing University of Aeronautics and Astronautics (NUAA), and deputy director of State Key Laboratory of Mechanics and Control of Mechanical Structures, China.

Dr. Hu was a research engineer at the R&D Center of NEC-Tokin, Sendai, Japan, from Nov. 1997 to Feb. 1999; research fellow and postdoctoral fellow at Hong Kong Polytechnic University, Hong Kong, China, from 1999 to 2001; assistant professor at Nanyang Technological University, Singapore, from 2001 to 2005; and associate professor at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2005 to 2010. His present research interest includes ultrasonic manipulators and

actuators, piezoelectric transducers and transformers, physical effects of ultrasound, wireless drive of piezoelectric components, energy harvesting from oscillation, and other novel utilization of vibration. He is a senior member of IEEE, and the Editorial Board Member of three international journals. Dr. Hu won the Paper Prize from the Institute of Electronics, Information and Communication Engineers (Japan) as the first author in 1998, and was awarded the title of valued reviewer of *Sensors and Actuators A: Physical and Ultrasonics*. He has given eight invited talks at international conferences, and is the honorary chairman of IWPMA 2011, held in USA. He is the author and co-author of more than 200 papers and disclosed patents, including more than 70 full papers published in SCI journals, and his research work in ultrasonic micro/nano manipulations has been highlighted by 7 international scientific media. He is also the author of monograph book “Ultrasonic Micro/Nano Manipulations” (2014, World Scientific, Singapore).

# The Gravity Control Experiments: Sensors, Equipment, Results

Vitaly O. Groppen

**Abstract** — The objective of this paper is to optimize the parameters of deployed capacitors used in the gravity control experiments as sensors. Construction and capacity of a deployed capacitor as well as applied voltage are the tools of this optimization. Used mathematical model is based on the idea of substitution of energy distributed in the neighborhood above the upper surface of the plate-deployed capacitor by the material point with equivalent mass: force of the gravitational interaction of the plate with this point has opposite direction to the force of gravitational interaction of this plate with the Earth thus reducing this force. Results of experiments with different sensors and scales allow us to select effective equipment and combination of capacity and voltage, which should have a deployed capacitor.

**Keywords** — deployed capacitor, experimental verification,<sup>1</sup> gravity control, high voltage.

## I. INTRODUCTION

The first model of the gravitational interaction forces was proposed by Sir Isaac Newton in 1667 [1]. About 250 years later, in 1915, Albert Einstein demonstrated a new theory of gravitation based on the Theory of Relativity [2]. In 1921 Townsend Brown discovered movement of physical objects under the influence of high voltage [3], but this effect cannot be considered as control of gravitational forces because this phenomenon is known to be caused by ionization of air near acute and sharp edges. The experiments described below are a continuation of the experiments presented in [4]-[6]. Their objective is to refine the parameters of used samples that enhance the lifting force. They are also based on the use of high voltage and charged deployed capacitors (Fig. 1) for gravity control. These experiments are based on the model using substitution of the energy distributed in the neighborhood above the upper surface of the deployed capacitor by the material point with equivalent mass: force of the gravitational interaction of the plate with this point is directed opposite to the direction of the force of gravitational interaction of this plate with the Earth therefore reducing the capacitor's weight (Fig. 2). As it is shown in [4], [5], such a weight reduction is proportional to the energy stored by a deployed capacitor.

This work is supported by the Grant # 262 of the Ministry of Education and Science of the Russian Federation.

V. O. Groppen is with the North-Caucasian Institute of Mining and Metallurgy (State Technological University), Nikolaev str. 44, Vladikavkaz 362021, North Ossetia, RUSSIA, phone: +78672407107, fax: +78672407203, e-mail: [groppen@mail.ru](mailto:groppen@mail.ru)

However, there are two opposite ways of increasing this energy. One of them is in increasing of capacitance of a deployed capacitor and, consequently, in reducing of distance between electrodes. To prevent the electric breakdown, the latter results in decreasing of voltage applied to a capacitor. Another way is to increase the voltage applied to the capacitor's plates, which entails an increase in the distance between them and, as a consequence, reduction of capacitance of a capacitor. Below we analyze the efficiency of both approaches.

## II. MAIN PRINCIPLES

Electrodes of used in experiments capacitors are designed as metal strips on a dielectric substrate forming thus deployed capacitor so that its' stored energy is distributed above the upper surface of a horizontally positioned capacitor (see Fig. 2a and Fig. 3a below).

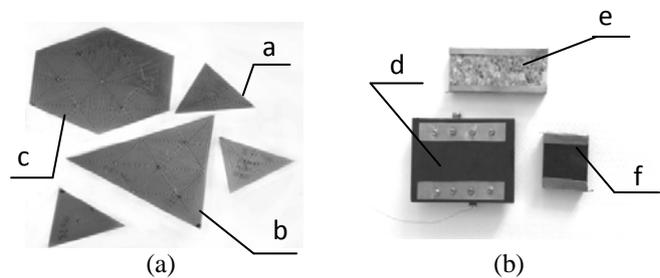


Fig 1. Deployed capacitors on fiberglass (a) and granite (b) substrate used during the experiments (top view)

The energy  $E_i$  of each  $i$ -th charged capacitor is equal to:

$$\forall i : E_i = \frac{C_i U^2}{2}, \quad (1)$$

where " $C_i$ " is its' capacity, " $U$ " - power supply voltage.

The mass of this energy is determined as follows:

$$\forall i : m(E_i) = \frac{C_i U^2}{2c^2}, \quad (2)$$

where  $c$  - velocity of light.

Below we suppose that:

- each capacitor is disposed horizontally, so that the electrodes are on its' upper surface;
- distributed above the upper surface of this capacitor

energy  $E_i$  is replaced by the equivalent body D, whose mass is determined by the expression (2).

Thus the force  $F_i$  of the gravitational interaction between the  $i$ -th capacitor and body D has a direction opposite to the force  $F_e$  of gravitational interaction between this plate and the Earth (Fig. 2).

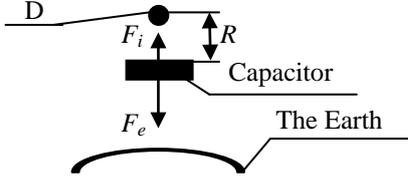


Fig.2. The forces of interaction the body D,  $i$ -th plate and Earth.

This lifting force  $F_i$  value in accordance with the Newton Law of gravity and equation (2), for any sample at Fig. 1 is determined as follows:

$$\forall i: F_i = \gamma \frac{m_i C_i U^2}{2R^2 c^2}, \quad (3)$$

where  $\gamma$  - gravitational constant,  $c$  - velocity of light,  $R$  - the shortest distance between the corresponding body D point and the surface of  $i$ -th capacitor.

Denoting  $F_e^0$  the weight of a plate before experiment

whereas  $F_e^1$  - its weight during experiment when the electrodes on its surface are applied to voltage equal to  $U$ , it is easy to determine lifting force value:

$$F_i = F_e^0 - F_e^1. \quad (4)$$

Fixing during each experiment all components of the equation (3) except distance  $R$ , the latter for  $i$ -th capacitor during  $j$ -th experiment can be determined as:

$$\forall i, R_i(U_j) = \frac{U_j}{c} \sqrt{\gamma \frac{m_i C_i}{2F_{i,j}}}. \quad (5)$$

Thus value  $R_i$  for each  $i$ -th sample may be determined as the arithmetic average of  $R_i(U_j)$ :

$$\forall i: R_i = \frac{1}{j_{\max}} \sum_{j=1}^{j_{\max}} R_i(U_j). \quad (6)$$

If we denote the subset of indices of capacitors having the same energy by the symbol "I", the best will be the  $k$ -th capacitor, which satisfies the following condition:

$$R_k = \min_{i \in I} R_i. \quad (7)$$

Since this distance, as it is shown below, is small, as the unit of its measurement below is used Fermi (Fm): 1Fm =  $10^{-15}$  m.

### III. EQUIPMENT, SAMPLES AND RESULTS OF EXPERIMENTS

As noted above, during the experiments were used two groups of samples: the first one was made in an effort to maximize the energy of charged capacitor via its' maximum capacity and simultaneously to minimize its' weight, whereas in the second group for the same goal we tried to maximize the capacitor's weight and the voltage applied to the capacitor, which does not lead to the fixed leakage current. The latter restriction was necessary to minimize the lifting force of Biefeld-Brown effect [3].

#### A. The first series of experiments

Geometry of electrodes of the first group samples used in the first series of experiments is shown below in Figure 3, whereas their main parameters - in Table 1.

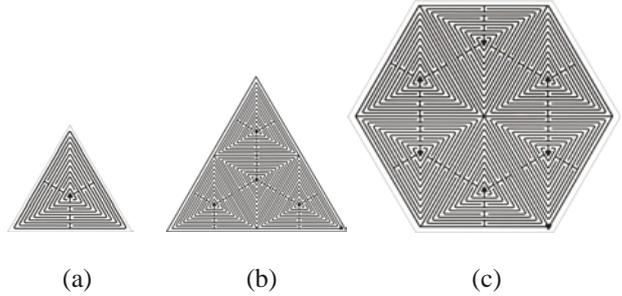


Fig. 3. Geometry of electrodes in the first group samples

It is easy to see that "b" and "c" samples (Fig. 3) consist of four ("b") and six ("c") triangles type "a". On the lamsan layer of "a" sample (Fig. 3, a) were fixed 12 copper nested equilateral triangles creating two groups of copper electrodes with width of these electrodes equal to 1 mm. and distance between the neighbor electrodes equal to 1.213 mm (see Table I). One group included six even triangles, another - six odd triangles and triangles in each group were connected electrically. During experiments center electrodes of all the "a"- triangles belonging to each "b" or "c" sample were connected electrically resulting in the parallel connection of corresponding capacitors.

In the experiments we used:

- a) the high voltage power supply IVNR-20/10, guarantying voltage range 1 – 20 kV, power 200 wt. (Fig. 4(a), 1);
- b) precise electronic scale AV-60/01-S which precision is equal to 0.0001 g, maximum weight – 60 g., the settling time of weighting mode – about 10 minutes (Fig. 4(a), 2);
- c) digital display of the electronic scale AV-60/01-S (Fig. 4(a), 3).

Table I • as shown in [ 4 ], any prolonged exposure of different samples based on fibre glass with lamsan cover to high voltage leads to its' electrical breakdown.

| № | Parameter name                  | Labels of samples in Fig. 1   |       |       | Units |
|---|---------------------------------|-------------------------------|-------|-------|-------|
|   |                                 | a                             | b     | c     |       |
| 1 | 2                               | 3                             | 4     | 5     | 6     |
| 1 | The length of one side          | 0.098                         | 0.194 | 0.098 | m.    |
| 2 | Weight                          | 5.0 ± 2.2                     | 19.7  | 33.23 | gm.   |
| 3 | Thickness                       | 0.6                           | 0.6   | 0.6   | mm.   |
| 4 | Distance between the electrodes | 1.213                         | 1.213 | 1.213 | mm.   |
| 5 | Width of the electrodes         | 1.0                           | 1.0   | 1.0   | mm.   |
| 6 | Capacity of the sample          | 33                            | 121   | 174   | pF    |
| 7 | Material of the basis           | Fiber glass with lamsan cover |       |       | -     |
| 8 | Material of the electrodes      | Copper                        |       |       | -     |

The parameters of deployed capacitors - samples presented at Fig. 3 and Fig. 1(a)

To minimize the errors indicated above, during the second series of experiments were used the other samples and equipment.

*B. The second series of experiments*

Within the second series of experiments we used:

- a) new samples with better resistance to electrical breakdown made of granite with two spaced apart parallel copper strips, attached to the top of each granite rectangle (Fig. 1(b));
- b) instead of precise electronic scale AV-60/01-S new precision mechanical balance AB-200 with maximum weight equal to 200 gram and precision equal to 0.001 g (Fig. 5(a)), which is not exposed to electromagnetic radiation.

The weight, capacity and geometrical parameters of the samples “d”, “e” and “f” shown at Fig. 1(b) are presented below in Table II.

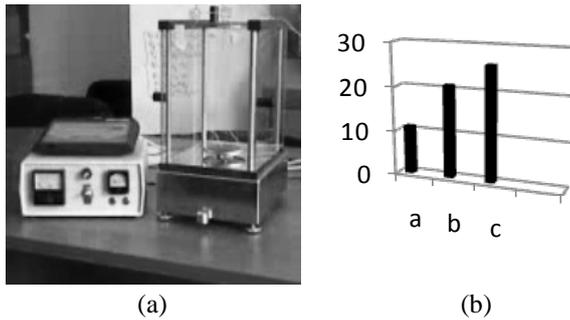


Fig. 4. Equipment used in the first series of gravity control experiments (a) and corresponding diagram (b) of distances  $R_i$  ( $i \in \{a, b, c\}$ ) values determined according to (6) for samples “a”, “b”, “c” (Table I).

The experimental data reflecting dependences of lifting forces on voltage for each sample – capacitor are presented in Appendix 1 below. There seems to be three typical sources of weight value mistakes during the first series of experiments:

- due to the proximity of the electrodes in the samples of the first series of experiments for a voltage greater than 3.5 kV have been substantial leakage currents, indicating the impact of the Biefeld - Brown effect on the weight of a sample.
- experiments for direct weight measurement of samples under high voltage resulted in direct interaction of electronic circuit of the scale and its sensor with the electric field of a sample often resulting in distortions in indications of weight by the scale and even in blocking the electronics of the scale;

Table II

| № | Parameter name                  | Labels of samples in Fig. 3a |          |       | Units            |
|---|---------------------------------|------------------------------|----------|-------|------------------|
|   |                                 | d                            | e        | f     |                  |
| 1 | 2                               | 3                            | 4        | 5     | 6                |
| 1 | Upper surface area              | 0.0063                       | 0.003072 | 0.001 | m <sup>2</sup> . |
| 2 | Total surface area              | 0.01586                      | 0.008704 | 0.004 | m <sup>2</sup> . |
| 3 | Weight                          | 211.0                        | 85.16    | 55.07 | g.               |
| 4 | Thickness                       | 10.0                         | 10.0     | 10.0  | mm.              |
| 5 | Distance between the electrodes | 30.0                         | 22.0     | 26.0  | mm.              |
| 6 | Width of the electrodes         | 11.0                         | 5.0      | 7.0   | mm.              |
| 7 | Capacity                        | 6.166                        | 2.9      | 1.6   | pF               |
| 8 | Material of the plate basis     | Granite                      |          |       | -                |

The parameters of deployed capacitors presented at Fig. 1(b)

Voltage and corresponding change of weight for each sample of the second group are presented in the Appendix 2, whereas diagram of  $R_i$ ,  $i \in \{d, e, f\}$  distances determined according to (6) for samples “d”, “e”, “f” (Table II) is presented at Figure 5(b). Samples of this diagram are ordered by increasing of their capacity.

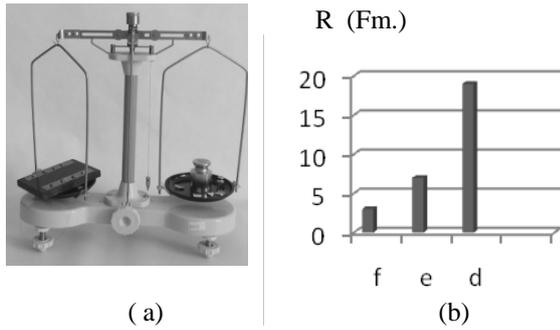


Figure 5. Mechanical scale AB-200 (a) and corresponding diagram (b) of distances  $R_i$  ( $i \in \{d, e, f\}$ ) values determined according to (6) for samples “d”, “e”, “f” (see Table II above).

IV. CONCLUSIONS

Using (6) and (7) and comparing the diagrams 4b and 5b for cases, where the energies gained in different capacitors are close, it is easy to see the benefits of the second approach: in the energy range  $2 \cdot 10^{-4} < E < 8 \cdot 10^{-4}$  (j) the following inequality holds:

$$\min_{i \in I_2} R_i \leq \min_{j \in I_1} R_j, \tag{8}$$

where  $I_1$  – the set of samples used in the first series of experiments;  $I_2$  – the set of samples used during the second series of experiments.

In other words, the above experimental results allow us the following conclusions concerning the future experiments on gravity control:

1. In the future experiments operating voltage of sensors - capacitors should be maximized.
2. It is preferred to use in experiments the precise mechanical scales, whose readings are independent of the electromagnetic fields.
3. It is necessary to create such a sensor which would minimize the value of R, at the same time withstanding high voltage U.
4. The range of the voltage applied during the experiments to the sensors should be expanded.
5. To exclude the influence on the readings of the balance of the charged particles in the air, the experiments should be repeated in an airless environment.

APPENDIX 1

Voltage and corresponding lifting forces  $F_i$ ,  $i \in \{a, b, c\}$ , for the samples “a”, “b” and “c” used during the first series of experiments (see Fig. 1(a)) are presented in Table III below:

Table III

| # | U (kV) | $F_a$ (N)                | $F_b$ (N)                | $F_c$ (N)                |
|---|--------|--------------------------|--------------------------|--------------------------|
| 1 | 2      | 3                        | 4                        | 5                        |
| 1 | 2.0    | $7.84532 \cdot 10^{-6}$  | $25.49729 \cdot 10^{-6}$ | $11.76798 \cdot 10^{-6}$ |
| 2 | 2.5    | $15.69064 \cdot 10^{-6}$ | $10.78732 \cdot 10^{-6}$ | $20.59397 \cdot 10^{-6}$ |
| 3 | 3.0    | $6.374323 \cdot 10^{-6}$ | $13.72931 \cdot 10^{-6}$ | $28.43929 \cdot 10^{-6}$ |
| 4 | 3.5    | $3.92266 \cdot 10^{-6}$  | $73.54988 \cdot 10^{-6}$ | $32.36195 \cdot 10^{-6}$ |

APPENDIX 2

Voltage and corresponding lifting forces  $F_i$ ,  $i \in \{d, e, f\}$ , for the samples “d”, “e” and “f” used during the second series of experiments (see Fig. 1(b)) are presented in Tables IV and V below:

Voltage and corresponding lifting forces  $F_d$  and  $F_e$  for the samples “d” and “e” (see Fig. 1(b)) are presented in Table IV below:

Table IV

| #  | U (kV) | $F_d$ (N)                | $F_e$ (N)                |
|----|--------|--------------------------|--------------------------|
| 1  | 2      | 3                        | 4                        |
| 1  | 9.0    | -                        | $1.412158 \cdot 10^{-4}$ |
| 2  | 10.0   | -                        | $1.90249 \cdot 10^{-4}$  |
| 3  | 11.0   | -                        | $2.755669 \cdot 10^{-4}$ |
| 4  | 12.0   | -                        | $2.843929 \cdot 10^{-4}$ |
| 5  | 13.0   | $1.833844 \cdot 10^{-4}$ | $3.6873 \cdot 10^{-4}$   |
| 6  | 14.0   | $2.265336 \cdot 10^{-4}$ | $5.138685 \cdot 10^{-4}$ |
| 7  | 14.5   | $2.628182 \cdot 10^{-4}$ | -                        |
| 8  | 15.0   | $2.598762 \cdot 10^{-4}$ | $5.295592 \cdot 10^{-4}$ |
| 9  | 15.5   | $3.138128 \cdot 10^{-4}$ | -                        |
| 10 | 16.0   | $3.854014 \cdot 10^{-4}$ | $7.404021 \cdot 10^{-4}$ |
| 11 | 17.0   | -                        | $7.58054 \cdot 10^{-4}$  |
| 12 | 18.0   | -                        | $8.482752 \cdot 10^{-4}$ |
| 13 | 19.0   | -                        | $10.86577 \cdot 10^{-4}$ |
| 14 | 20.0   | -                        | $11.33649 \cdot 10^{-4}$ |

Voltage and corresponding change of lifting force  $F_f$  for the sample “f” (Fig. 1(b), f) are presented in Table V:

Table V

| # | U<br>(kV) | $F_f$ (N)                 |
|---|-----------|---------------------------|
| 1 | 2         | 3                         |
| 1 | 2.5       | $1.137571 \cdot 10^{-4}$  |
| 2 | 3.0       | $1.274865 \cdot 10^{-4}$  |
| 3 | 4.0       | $2.186883 \cdot 10^{-4}$  |
| 4 | 5.0       | $1.549451 \cdot 10^{-4}$  |
| 5 | 7.5       | $2.43205 \cdot 10^{-4}$   |
| 6 | 10.0      | $1.78481 \cdot 10^{-4}$   |
| 7 | 14.0      | $2.16727 \cdot 10^{-4}$   |
| 8 | 15.0      | $0.8825985 \cdot 10^{-4}$ |

## REFERENCES

- [1] Newton. The mathematical principles of natural knowledge, 1667.
- [2] A. Einstein. The theory of relativity. *Die Physik*, Under reduction of E. Lechner, Leipzig, V. 3, 1915, pp. 703 – 713, (in German)
- [3] D. R. Buehler. Exploratory Research on the Phenomenon of the Movement of High Voltage Capacitors, *Journal of Space Mixing*, April 2004, vol. 2, pp. 1-22,
- [4] V. O. Groppen, Gravity control: modeling and experiments. *Proceedings of the 2014 International Conference on Energy, Environment, Ecosystems and Development II (EEED'14)*, Prague, Czech Republic, April 2 – 4, 2014, pp. 15 – 17.
- [5] V. O. Groppen . Control of the Forces of Gravity: Modeling and Experimental Verification *WSEAS Transactions on Applied and Theoretical Mechanics*, ISSN / E-ISSN: 1991-8747 / 2224-3429, Volume 9, 2014, Art. #19, pp. 215-221
- [6] V. O. Groppen . Manifestations of Measurement Standards Variability in the Universe Modeling, *Lambert Academic Publishing*, Saarbrücken, Germany, 2013, 76p.

# Influence of the Applied Electric Field on the Growth of an Electrical Discharge

L. Zeghichi, L. Mokhnache, and M. Djebabra

**Abstract**— This paper describes formulation of a Monte Carlo model, which is capable of describing electron dynamics. At higher fields, charged particles may gain sufficient energy between collisions to cause ionization on impact with neutral molecules. Ionization by electron impact under strong electric field is the most important process leading to breakdown of gases.

The avalanche growth is simulated by tracing individual paths of charged particles. When the electron multiplication is large the difference in mobility of electrons and positive ions introduces a space charge field which distorts the applied field. and the effect of space charge is included by solving the Poisson equation. The simulation is carried out in O<sub>2</sub> gas under the effect of uniform electrical fields. The streamer breakdown criterion for the different applied uniform fields is examined.

**Keywords**—Electrical Breakdown, Collision probability, Electrical discharge, Poisson's equation, Monte Carlo Simulation.

## I. INTRODUCTION

THE term “discharge” was applied to any flow of electric current through gas, and to any process of ionization of the gas under the effect of an applied electric field. The modern field of gas discharge physics is thus occupied with processes connected with electric currents in gases and with generating and maintaining the ability of a gas to conduct electricity [1].

Gases are important in the field of high voltage engineering. They are used mainly for the insulation and prevention of electrical breakdown in high voltage circuits and transmission lines [2].

The type of discharge is determined by the various physical conditions of gases, namely, pressure, temperature, electrode field configuration, nature of electrode surfaces, and the availability of initial conducting particles are known to govern the ionization processes [3-7]. The breakdown voltage of a given gap depends on the gas parameters such as the ionization coefficient ( $\alpha$ ), the attachment coefficient ( $\eta$ ), the recombination coefficient ( $\beta$ ), and the Townsend second ionization coefficient ( $\gamma$ ), which in turn are functions of the electric field and of such factors [8].

The Monte Carlo Simulation (MCS) is used to describe the

different phenomena in the discharge process such as: the elastic and inelastic (ionization, attachment, and excitation) processes.

The MCS consists on tracking the electrons' trajectories; it is based on the mean free path or the mean free flight time. The physical parameters of the molecules that compose the studied gas: Collision cross section, Collision probability, Collision energy (elastic, attachment, excitation and ionization), are used to model the gas discharge. By using sampling laws, we obtain the parameters of electronic avalanches' development and growth: mean energy values, ionization and attachment coefficients.

In this paper we use the simulation results to verify the breakdown criterion and to find the solution of Poisson equation for the space charge field which is the main parameter producing the discharge. The effect of the applied electric field is investigated.

## II. SIMULATION METHOD

MCS is a stochastic method; it applies to problems with absolutely no probabilistic content in addition to those with intrinsic probabilistic structure. This method is based on a set of stochastic algorithms providing the approximation of numerical quantities by performing statistical sampling experiments on a computer. Pseudorandom numbers are used to describe the development of the real system in question.

The MCS has come to be known as the only approach capable of providing useful imitating tool for the electron's motion in gas discharge physics.

Monte Carlo experiment generates randomly a group of trial electrons. The application of the constant step MCS version for the study of electron's motion, under the effect of the electric field, requires the evaluation process, after experiencing energy loss and gain, of the different parameters taking into account the different processes of atomic collisions (elastic or inelastic).

### A. Collisions' Treatment

We have adopted a free flight time approach; the electron mean free flight time between two successive collisions is determined by the electron collision total cross section  $Q(\epsilon)$  as:

$$T_m = \frac{1}{N \cdot Q(\epsilon) \cdot v(\epsilon)} \quad (1)$$

L. Zeghichi, is now with the Department of Physics, Ouargla University, P.O Box.511, OUARGLA 30 000, Algeria(e-mail: zeghichi.leyla@univ-ouargla.dz).

L. Mokhnache is with the Electrical Engineering Department, Batna University, Batna, Algeria(e-mail: lmokhnache@yahoo.fr).

M. Djebabra is with the Institute of Health and Safety, Batna University, Batna, Algeria(e-mail: mebarek\_djebabra@yahoo.fr).

where:  $v(\varepsilon)$  is the drift velocity of electrons and  $N$  the gas number density.

The free flight time is divided into a number of smaller elements according to:

$$dt = \frac{Tm0}{K} \quad (2)$$

where:  $K$  is a sufficiently large integer.

The collision probability  $P_1$ , that follows the Poisson's distribution, is given by:

$$P_1 = 1 - \exp\left(-\frac{dt}{Tm}\right) \quad (3)$$

The interval  $[0, P_1]$  is divided into segments of lengths that correspond to the probabilities of different types of collision after increasing scheduling of these probabilities.

The remaining portion of the interval  $[0, 1]$  is for the case where no collision is possible.

The electron energy is described as follows:

For the elastic collision the energy is given by:

$$\varepsilon_1 = \left(1 - 2\frac{M}{m}\cos(\delta)\right)\varepsilon_0 \quad (4)$$

where:  $\delta$  is the scattering angle of the electron after the collision,  $m$  and  $M$  are, respectively, the masse of electron and an  $O_2$  molecule and  $\varepsilon_0$  is the electron's energy before collision.

For the processes (attachment, excitation and ionization), the onset energy "los" of the process is subtracted from the electron energy:

For an attachment of the electron, all its energy is to be lost, and therefore it is lost in the swarm.

$$\varepsilon_1 = 0 \quad (5)$$

For an exciting process of a molecule to a higher stat (different rotations, vibrations and electronic excited stats), the energy of the electron is reduced with the energy needed to excite the molecule and the resulting energy is given by:

$$\varepsilon_1(m) = \varepsilon_0(m) - los \quad (6)$$

And for an ionizing process, the remaining energy is shared between the, primary and ejected, electrons with the ratios  $R$  and  $(R-1)$  as:

$$\begin{aligned} \varepsilon_{primary} &= R(\varepsilon_0 - los) \\ \varepsilon_{ejected} &= (R-1)(\varepsilon_0 - los) \end{aligned} \quad (7)$$

where:  $R$  is a uniform random number between zero and unity.

### B. Implementation

At time  $t = 0$ , the initial electrons are emitted from the

cathode according to a cosine distribution. The energy gain of the electrons in a small time interval  $dt$  is governed by the equation of motion. The occurrence of collision between an electron and a gas molecule and its kind are determined by comparison of the collision probability  $P_1$  with computer generated random numbers  $R$ . The nature of the collision is determined in the following way:

The total number of electrons in the gap increases over many orders of magnitude. To limit the number of simulation particles, a statistical subroutine is introduced, when the total number of simulation particles exceeds the maximum number  $N_{max}$  permitted, to choose a new group of larger particles to represent the old larger group of smaller particles.

## III. SAMPLING LAWS

### A. Electron Swarm Parameters

The statistical treatment is carried out to obtain the physical parameters such as the attachment and the ionization coefficients. For that we exploit the key quantity  $\bar{z}$  which is given by the sampling formula below:

$$\bar{z} = \frac{1}{N} \sum_{i=1}^N z(i) \quad (8)$$

The ionization coefficient  $\alpha$

$$\alpha = \frac{\ln\left[\left(\frac{N_{pion}}{n_0}\right) + 1\right]}{\bar{z}} \quad (9)$$

The attachment coefficient  $\eta$

$$\left\{ \begin{aligned} \eta &= \frac{N_{nion}}{n_0} \frac{1}{\bar{z}} \quad (\text{if } \alpha = 0) \\ \eta &= \frac{N_{nion}}{N_{pion}} \alpha \end{aligned} \right. \quad (10)$$

where:  $n_0$  is the initial electrons number,  $N_{pion}$  and  $N_{nion}$  are, respectively, the number of positive ions et the number of negative ions (counters).

The mean energy of electrons  $\bar{\varepsilon}$  is known as

$$\bar{\varepsilon} = \frac{1}{N} \sum_{i=1}^N \varepsilon(i) \quad (11)$$

### B. Space Charge Field

To obtain the space charge field distribution, we use the distributions of the new electrons, positive and negative ions, which are produced by ionization and attachment processes, to resolve the Poisson equation.

IV. RESULTS AND DISCUSSION

In this paper we describe the development of an electrical discharge in O<sub>2</sub> by MCS in plane-plane geometry. The calculations are performed at gas pressures of 1 and 100torr. The cross section set of the O<sub>2</sub> molecule used is that referred in [9].

At  $t = 0$ , a number of electrons are released from the cathode with small energy 0.1 (eV).

The figures (Fig. 1, Fig. 2, Fig. 3) show, respectively, the variation with time of the ionization coefficient ( $\alpha$ ), the attachment coefficient ( $\eta$ ) and the mean kinetic energy; for Oxygen gas at a pressure of 1 (torr), a temperature of 293 (K)20(° C) with a gap length of 2 (cm) and an applied electric field of 10 (kV/m).

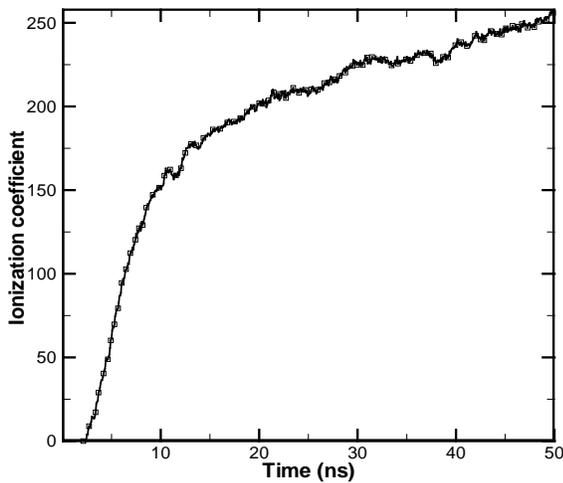


Fig. 1 Temporal variation of the ionization Coefficient at P=1torr and E0=10 kV/m.

At low pressure 1 (torr), the ionization coefficient increases with time, however, the attachment coefficient decreases but the Townsend breakdown criterion is not verified. At time  $t = 50$  (ns), the number of the total space charge is about 16307 (10065 electrons).

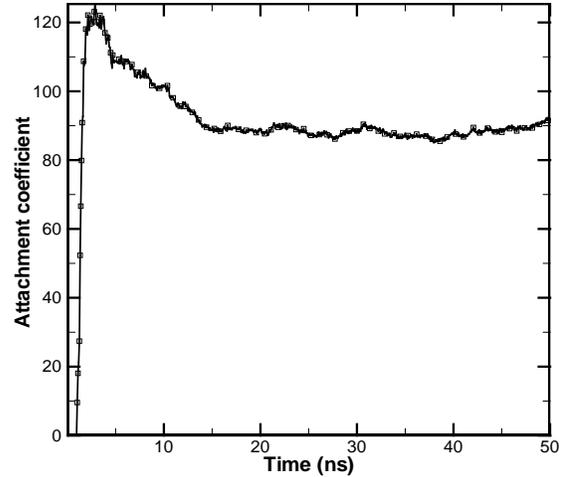


Fig. 2 Temporal variation of the attachment coefficient at P=1torr and E0=10 kV/m.

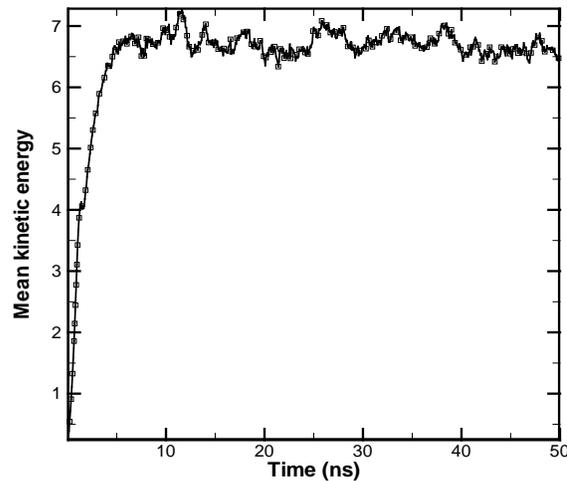


Fig. 3 Temporal variation of the mean kinetic energy at P=1torr and E0=10 kV/m.

The figure (Fig. 4) shows the spatial repartition of the space charge field at pressure of 1 (torr) under an applied electric field of 10 (kv/m) where we deduce that the space charge field is intense near the cathode but it is not sufficient to sustain the discharge. This result is in good agreement with the value given E. Kuffel [4].

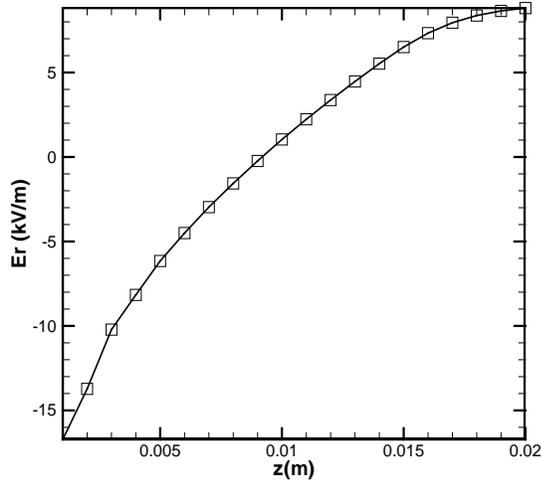


Fig. 4 Spatial charge field distribution at P=1 torr and E0=10 kV/m.

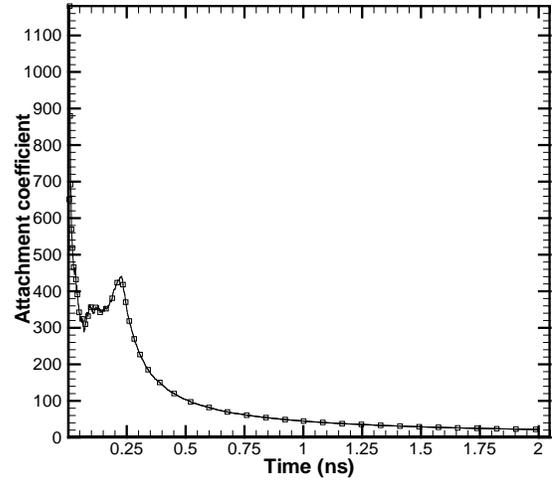


Fig.6 Temporal variation of the attachment coefficient at P=100 torr and E0=1000 kV/m.

The figures (Fig. 5 and Fig. 6) show, respectively, the variation with time of the ionization coefficient ( $\alpha$ ) and the attachment coefficient ( $\eta$ ); for Oxygen gas at a pressure of 100 (torrs), a temperature of 293 K (20° C) with a gap length of 2 (cm) and under an applied electric field E0=1000 (kV/m).

The figures (Fig. 7 and Fig. 8) show the distribution of the electric field Er due to space charge at times t = 2ns and t = 2.22ns respectively.

At time t= 2 ns, the total number of space charge is about 784434 and the streamer formation criterion is not verified.

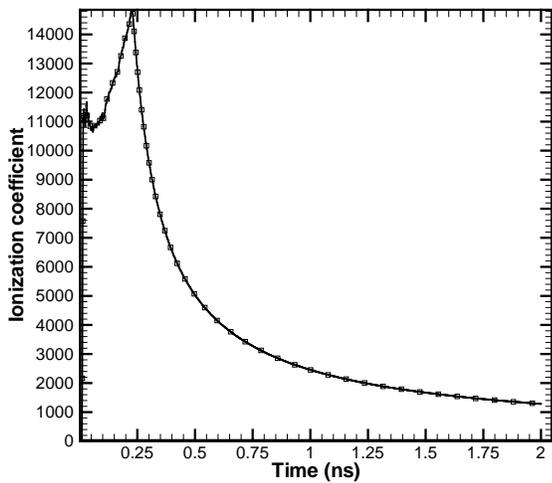


Fig. 5 Temporal variation of the ionization coefficient at P=100 (torr) and E0=1000 (kV/m).

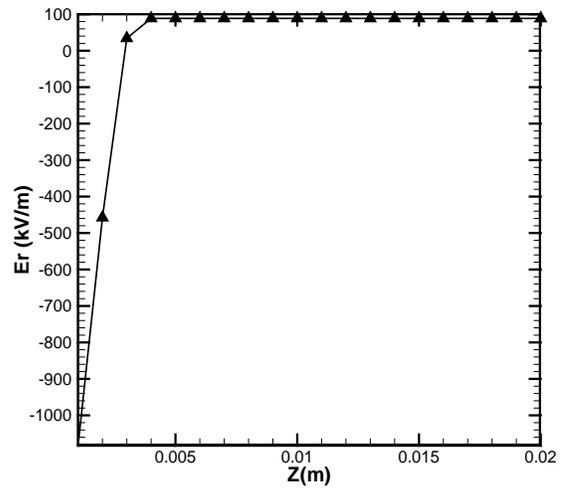


Fig. 9 Spatial distribution of the space charge field at P=100 (torr) and E0=7000 (kV/m).

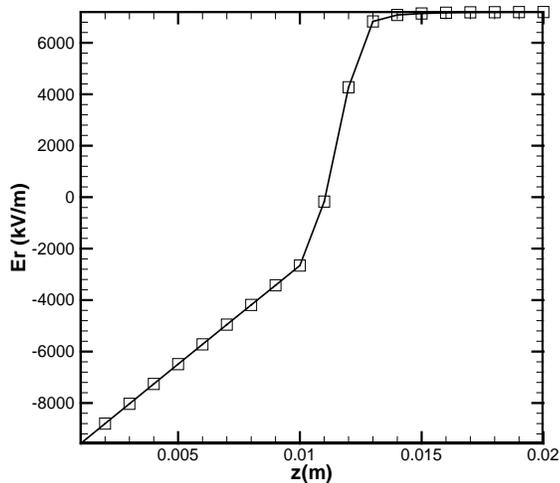


Fig. 10 Spatial charge field distribution at  $t=2.22\text{ns}$   $P=100$  torr and  $E_0=7000$  kV/m.

For Oxygen gas at a pressure of 100 (torr) and a temperature of 293 K (20° C) with a gap length of 2 (cm):

At time  $t=2.22$  ns, under an applied electric field  $E_0=7000$  (kV/m), the total number of space charge is of 9846099 and it is able to produce sufficiently strong electric field which allow secondary electrons due to photo-ionization to develop into the avalanche head. The streamer breakdown criterion  $\bar{\alpha}d=18.03$  is verified; so we can say, in this case, that the streamer breakdown is established

## V. CONCLUSION

In this paper we have used the Monte Carlo Simulation to describe the behavior a large number of electrons in uniform electric fields.

For reasons of simplification, our model is based only on

The simulation results give values for electrons mean energy, ionization and attachment coefficients as functions of time. When the voltage is sufficiently high, the ionization coefficient increases and the gas become conductor, and therefore the appearance of the electrical breakdown.

By means of the simulation results we have verified the breakdown criteria for different values of applied electric field (pressure and applied voltages) (Townsend for the low pressures and Streamer for the high pressures)

## REFERENCES

- [1] Y. R. Raizer, "Gas Discharge Physics", Berlin: Springer-Verlag, 1991.
- [2] M. M. Pejovic, G. S. Ristic and J. P. Karamarkovic "Electrical breakdown in low pressure gases", *J. Phys. D: Appl. Phys.* 35, pp. R91–R103, 2002.
- [3] J. M. Meek, J. D. Craggs, *Electrical Breakdown of Gases*, Oxford : Clarendon Press, 1953.
- [4] E. Kuffel, W. S. Zaengl and J. Kuffel: *High Voltage Engineering Fundamentals*, 2nd ed Butterworth-Heinemann, 2000, 534 p.
- [5] M. S. Naidu, *High Voltage Engineering*, 2nd ed. New York: Quebecor/Book Press, 1995.

- [6] G. R. Govinda Raju and J. Liu. "Simulation of Electrical Discharges in Gases–Nonuniform Electric Fields", *IEEE Transactions on Dielectrics and Electrical Insulation*, vol 2 (5), pp. 1016-1041, 1995.
- [7] G. G. Raju, *Dielectrics in Electric Fields*, Marcel Dekker, New York: CRC Press, 2003.
- [8] [8] G. R. Govinda Raju and J. Liu. "Simulation of Electrical Discharges in Gases– Uniform Electric Fields". *IEEE Transactions on Dielectrics and Electrical Insulation*, vol 2 (5), pp. 1004-1015, 1995.
- [9] [9] A. V. Phelps, Atomic & Molecular Physics. JILA NIST-CU website. [Online] 2005. Available:
- [10] [ftp://jila.colorado.edu/collision\\_data/electronneutral/ELECTRON.TXT](ftp://jila.colorado.edu/collision_data/electronneutral/ELECTRON.TXT)

**L. Zeghichi** Was born in Batna in 22 Jun 1983. She received the degree of graduation in physics of radiation from BatnaUniversity (Algeria) in 2006 and she obtained the Magister in physics of radiation (Laser and Plasma) from BatnaUniversity (Algeria) in 2010. The author's major field of study should be lower-cased.

She is assistant professor, since November 2011, at Ouargla University (Ageria).

# A novel flexible electrodynamic planar loudspeaker

Jium-Ming Lin, Ubadigha Chinweze Ukachukwu, and Cheng-Hung Lin

**Abstract**—This paper proposed a flexible electrodynamic planar loudspeaker (FEPL) with limited thickness ( $\ll 10$  mm). The structure is very simple such as a flexible thin film diaphragm (polyimide) electroplated traces of copper coil above a flexible magnetic placed in the bottom of cavity, thus forming a seamless integration of electromagnet actuation and planar flexible structure. The advantage of this design is that it can be used in flexible electronics or can be deployed on the surface of any object easily. Compared with an equivalent cone type loudspeaker, the performance of FEPL infinite baffle was found to be better at high frequencies but lags slightly in the low frequency range. Three additional cases of the FEPL were investigated, i.e. the FEPL with enclosure and no vent (Case 1), the FEPL with enclosure and vented around the side walls (Case 2), and the FEPL with enclosure and vented on the diaphragm (Case 3). In general, the FEPLs with enclosure showed a better performance than the infinite baffle one. However, for the FEPL with enclosure and no vent (Case 1), the performance is only better in the high frequency region, and the operational range is 1.3 to 20 kHz. The FEPLs with enclosure and vents around the side walls (Case 2) showed a better sensitivity extending towards the lower frequency region, the operational range is 60Hz to 20 kHz and with a minimum average SPL of 50 dB in the lower frequency region and 90 dB in the higher frequency region. On the other hand, the performance Case 3 is between those of Cases 1 and 2. To optimize the performance, this study made a detailed analysis on the thickness of cavity, magnet and coil, magnet configuration and polarization, and diaphragm dimension, thus contributing to the scarce literature in this area of study.

**Keywords**—Electromagnetic actuation, flexible substrate, magnetic flux density, planar loudspeaker, sound pressure level.

## I. INTRODUCTION

**F**LEXIBLE electronics has been a hot research and development topic in the electronics industry since the past few years. This is due to the rapid growth of flexible electronic technology [1]–[2]. The speaker is an important part of the electronics and has received many attentions [3]–[11]. Some progresses have been made towards the development of flexible

loudspeaker, and most of them have failed for commercial production. In loudspeaker design, many different actuation mechanisms are employed for electro-acoustic transduction, such as electromagnetic [4], [12]–[16], piezoelectric [5]–[6], electrostatic [17, 8], and electro-thermal [18]–[19] actuation mechanism. The flexible and transparent loudspeakers using piezoelectric actuation mechanism were developed [3, 5] by using PVDF as the piezoelectric polymer as radiator. Results of their study showed that the PVDF driven flexible loudspeaker were able to produce 70 dB and 80 dB SPL within a frequency range of 1 to 20 KHz and 400 Hz to 10 KHz, respectively. However, PVDF material is very expensive due to the complex production process [6]. They are of high frequency speakers and difficult to produce sound in low frequency range. Industrial Technology Research Institute (ITRI) Taiwan in 2009 filed a patent [17] of an electrostatic actuated ultrathin flexible loudspeaker, it was able to produce sound within 200 Hz -20 KHz, and thus can be operated in medium and high frequency [11]. They combined arrays of tiny, bendable speakers to produce speaker systems of almost any size using standard inkjet printing on substrate of paper or plastic and a thin metal. Though suffers the same fate as PDVF loudspeaker in producing low frequency sound. Furthermore, other industrial based developed flexible loudspeakers include Yamaha Corporation [9] and Warwick audio technologies [10] that developed an electrostatic directional flexible loudspeaker capable of producing sound only in a specified direction. Fujifilm [8] also developed an electro-acoustic film which also operates using electrostatic mechanism. The inspired work of Xiao *et al.* [19] brought to light a flexible, stretchable, transparent loudspeaker designed using carbon nano-tubes (CNT). It operates using the electro-thermal mechanism. But this loudspeaker has a major drawback for the lack of industrial process to create thin films of CNTs. Until now electromagnetic actuation mechanism has not been explored in designing flat flexible loudspeaker that can be used in pop up banners, portable exhibition stands and other uses that require flat flexible loudspeaker.

This study has explored the possibility to develop a flexible electrodynamic planar loudspeaker using a finite element method (FEM) approach. Over the years, electromagnetic actuation has been proven to be the most efficient actuation mechanism to generate sound pressure [4]–[13]. In general the electrodynamic loudspeakers generate sound through the interaction of a magnetic field, usually created by a rigid

This work was supported in part by National Science Council with the grants: NSC 101-2622-E-216-001-CC3, 101-2221-E-216-006-MY2, 101-2221-E-216-019-, and 102-2622-E-216-002-CC2.

Jium-Ming Lin is with Department of Communication Engineering, Chung-Hua University, Hsin-Chu, 30012 Taiwan, R. O. C. (corresponding author: 886-3-5186483; fax: 886-3-5186521; e-mail: [jmlin@chu.edu.tw](mailto:jmlin@chu.edu.tw)).

Ubadigha Chinweze Ukachukwu, was with Department of Mechanical Engineering, Chung-Hua University, Hsin-Chu, 30012 Taiwan, R. O. C. (e-mail: [b09306014@chu.edu.tw](mailto:b09306014@chu.edu.tw)).

Cheng-Hung Lin is with Ph. D. Program in Engineering Science, College of Engineering, Chung-Hua University, 30012 Taiwan, R. O. C. (e-mail: [b09306014@chu.edu.tw](mailto:b09306014@chu.edu.tw)).

permanent magnet (neodymium magnet), with a coil of wire carrying an audio current and attached to a diaphragm. The proposed design adopts the structure of a single ended planar loudspeaker which is an electrodynamic loudspeaker [7, 14]. Therefore, the challenge is how to utilize this actuation mechanism and structure to achieve a flexible loudspeaker. A typical structure of the proposed design is shown in Fig. 1. The core structure is the substitution of the conventional rigid magnet with a flexible magnet made of mixture of polymer and neodymium (NdFeB) or ferrite material. According to [21] it is obtainable to have flexible magnets that have up to 1.8 MgOe (2730 gauss) or more. This is relatively good for flexible loudspeaker application when all other militating factors are put to check as will be shown later. Besides, an ultrathin flexible polyimide film was adopted as the diaphragm with a copper coil electroplated on one side of its surface. The proposed flexible loudspeaker is expected to have a thickness less than 10 mm.

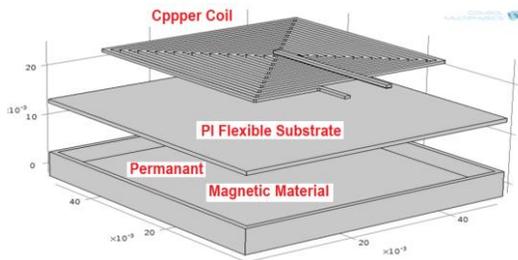


Fig. 1 Exploded view of the basic design of the proposed FEPL

The main focus of this study is to design an optimum structure that will yield the maximum efficiency and still retains the desired properties of the proposed loudspeaker and also analyze its mechano-acoustic and electrodynamic properties using FEM. The finite element simulation was done in two parts; first the FEPL was simulated considering it having an infinite baffle, this is to extract the essential parameters of the FEPL. Secondly, using the essential parameters the FEPL was modelled having an enclosure, with and without perforations. The pressure exerted inside the enclosure (behind the flexible substrate) was solved for and its effect on the SPL depicted. The results of the analysis showed that when supplied with a one volt audio signal, the proposed speaker is capable of producing 50 dB<sub>SPL</sub> (for FEPL to be as an infinite baffle) and 90 dB<sub>SPL</sub> (for FEPL to be with enclosure) measured from one meter distance, and have operating frequency ranges of 170 Hz to 20 KHz and 1.2 KHz to 20 KHz, respectively. Compared with an equivalent cone type loudspeaker, the performance of FEPL infinite baffle was found to be on par with the one of cone type loudspeaker at high frequencies but lags slightly in the low frequency range. To optimize the FEPL performance, this study made a detailed distribution analyses on the thickness of cavity, magnet and coil, magnet configuration and polarization, and diaphragm dimension, thus contributing to the scarce literature in this area of study. The paper is organized as follows: the first section is an introduction; the next part illustrates the FEPL structure configuration and optimization; Section 3 is results and discussions; the last part is a conclusion.

## II. FEPL STRUCTURE CONFIGURATION AND DESIGN

### A. Speaker Cavity Design

Considering the proposed structure of the FEPL the factors responsible for the performance includes the FEPL cavity (i.e., the distance of the coil from the magnet), the magnet polarization and the magnet arrangement adopted. Fig. 2 shows the 2-D cross-section view of FEPL. Fig. 3a shows the measured magnetic flux density (at 0.5 mm above the speaker surface) for four cases of speaker cavity height. As the cavity height increases, the magnetic flux density decreases according to the power law  $B^n$  where  $n = 1.0033, 1.00395, 1.0046$  and  $1.00525$  respectively for cavity height as 2 mm, 3 mm, 4 mm and 5 mm.  $n$  is found to be making a constant progression for every 1mm cavity height increment thus making the magnetic flux diminish rapidly as the magnet moves farther away. However, the larger the size of speaker enclosure the better chances of having a low frequency sound production if the resonant frequencies are well controlled [24]. A trade-off was employed here considering the speaker size and the diaphragm excursion. While 2 mm cavity height might be more appealing considering  $B$ , it is not suitable for a large excursion of the flexible substrate and mitigates the chances of achieving a low frequency operation of the FEPL. Thus, 4mm was chosen to be the suitable height for this design.

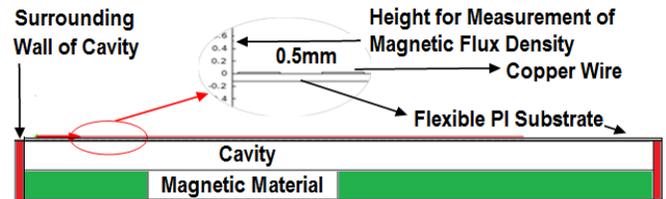


Fig. 2 2-D cross-section view of FEPL

### B. Magnet Thickness Design

Fig. 3b shows the magnitude of magnetic flux for different magnet thickness measured at a given distance (0.5 mm above coil surface) as depicted in Fig. 2. The magnitude of magnetic flux increases proportionally to the magnet thickness. However, while a 2 mm magnet thickness might not create enough magnetic fields, and the increase up to 5 mm would not be satisfactory because both the volume and weight are too larger. Thus, the thickness of magnet was chosen to be 4mm for the cost, packaging and sound quality trade-offs.

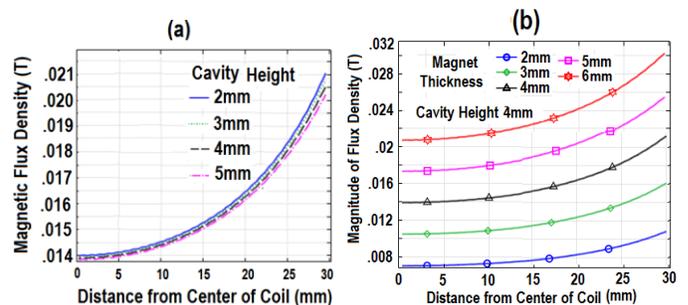


Fig.3 Magnitude of B for different (a) cavity height, and (b) magnet thickness

### C. Magnet Configuration and Polarization Design

Two basic modes of magnet polarization as respectively shown in Figs. 4a (vertical) and 4b (horizontal) were adopted. Based on the physics of electromagnetism, the component of a magnetic flux responsible for actuating a vertical force in perpendicular with the flat surface of a coil is the radial flux,  $B_r$ , of a planar magnet irrespective of its polarization. However, because the loudspeaker under investigation is of square shape, two magnetic flux components are responsible for the vertical force actuation; the x-component  $B_x$ , and the y-component  $B_y$ . The dimensions of the planar magnets used for investigation are as shown in Figs. 4a and 4b. Figs. 5 and 6 show the magnetic flux distributions and magnitudes in the x and y axes of a single magnet which is vertically polarized or horizontally polarized in the x-axis. Note that the flux density is concentrated on the edges of the magnet and less on the center by using the vertical type in either x- or y-axis (Figs. 5b, 5c and Fig. 6); while the flux density is concentrated on the edges of the magnet in the y-axis (Fig. 5f and Fig. 6) and less on the center in the x-axis (Fig. 5e and Fig. 6) by using the magnet horizontally polarized in the x-axis. The asymmetric nature of the horizontally polarized magnet will cause distortions in the diaphragm and lead to poor sound quality. So the vertically polarized magnet was chosen for the FEPL design.

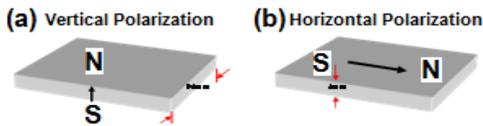


Fig.4 Polarization of B

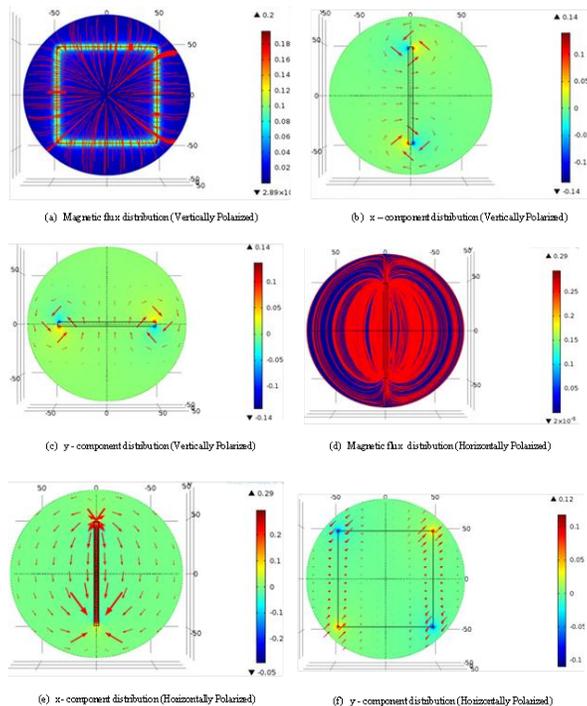


Fig. 5 Magnet flux density distribution for vertically polarized (a), (b), (c) and horizontally polarized (d), (e), (f)

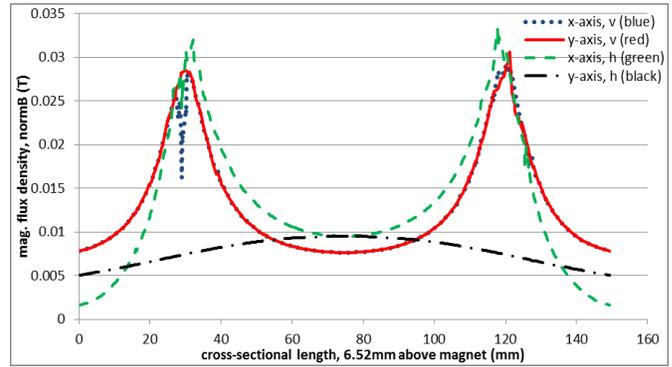


Fig. 6 Magnitudes of B in x and y axes for polarized in vertical, and horizontal planes, respectively

### D. Diaphragm Dimension Design

Polyimide was chosen because its temperature is endurable to 200°C and light weight. The density, coefficient of thermal expansion and thickness of the diaphragm are as 1430 kg/m<sup>3</sup>, 5.5×10<sup>-5</sup>/K and 0.122 mm, respectively. Finally, the performance of SPL and the dimension of the diaphragm are to trade off. Fig. 7 shows the SPL of the speaker for four surface areas (S) as 60 mm × 60 mm, 70 mm × 70 mm, 80 mm × 80 mm, and 90 mm × 90 mm. Note that as S increases the SPL slightly decreases and has more distortion within the operational frequency range. When S is 60 mm × 60 mm (blue), then the SPL is maximum but with a large resonance peak at 1500Hz. In comparison, the case of 70 mm × 70 mm has a more flat SPL than the one of 60 mm × 60 mm. Given this, the surface area with 70 mm × 70 mm is found to be the better one of diaphragm dimension, and the calculated mass is 0.855g.

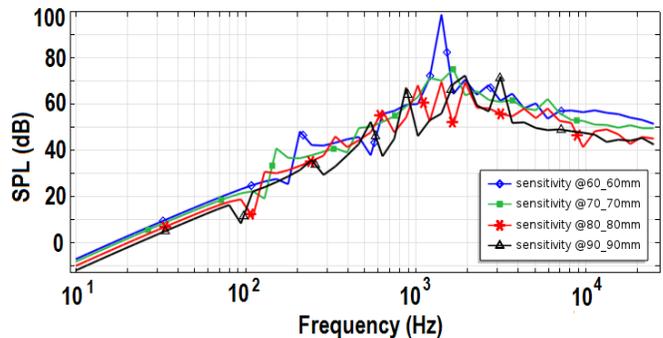


Fig. 7 SPL for four cases of diaphragm surface area (S)

### E. Coil Thickness Design

The turn and pitch of the coil are set as 30 turns and 1mm, respectively. Copper was chosen as the coil material for good electric conductivity. Fig. 8 shows that the SPL increases as the thickness (T) of the coil increases, which holds true, because the wire cross-sectional area is a function of the current density. However, at higher frequencies the trend changes as the effect of the coil weight becomes noticeable. For T > 0.1 mm, additional coil thickness starts playing a more diminishing role on the SPL in the high frequency region. Because as the coil thickness increases, the coil mass starts contributing negatively to the speaker SPL.

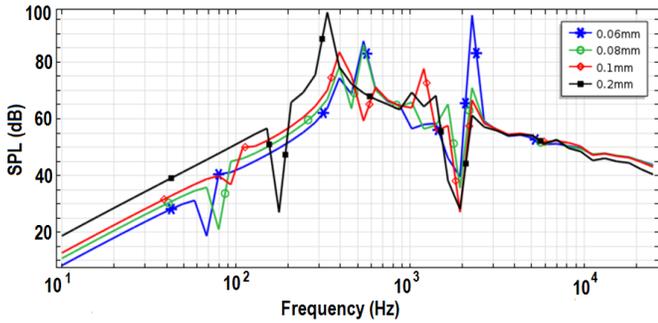


Fig. 8 SPL for four cases of coil thickness (T)

The better case of T is 0.1 mm. If  $L$  and  $W$  are the length and width of the first coil turn respectively,  $p$  is the pitch of the coil,  $l$  is the total length of coil, and  $N0$  is the number of coil turn. By (1) the total coil length was calculated to be 3.78 m. Consequently, the mass of the coil was calculated to be 1.6443 g given that copper has a density of  $8700 \text{ kg/m}^3$  and the coil is a square coil with evenly distributed pitch.

$$2N0(L + W) + 8p \sum_{i=1}^{N0} (N0 - 1) = l \quad (1)$$

### III. RESULTS AND DISCUSSIONS

Fig. 9 shows the SPLs for the FEPL to be as an infinite baffle and with enclosure. The operational frequency of the FEPL spans from 170 Hz to 20 kHz as an infinite baffle; this is considered the region where SPL curve is more flat and also to avoid the mechanical resonance which occurred at 157 Hz as shown in Fig. 10 for the maximum displacement plot of diaphragm over frequency. The operational frequency produces minimum and maximum SPLs of 40 dB and 110 dB respectively. Fig. 9 also shows the SPL of an equivalent electrodynamic cone speaker (conventional speaker), in which the parameters such as voltage applied, number of coil turns, coil cross-sectional area and the magnet remanent flux density of the speakers were set to equal values. But the radiating surface area of the conventional speaker is approx. 97% greater than that of the FEPL. The SPLs were measured from the same distance. Note that the conventional speaker has a better performance in the lower frequency range and produces a more stable sound within this region; because it can undergo a more space for linear motion (piston motion) within this frequency. While in the higher frequency region, the FEPL is seen to have a more stable sound.

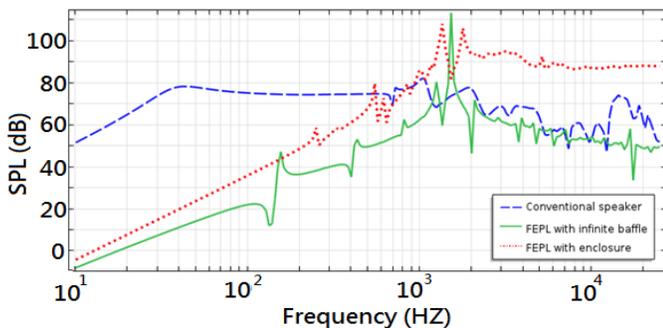


Fig. 9 SPLs of FEPL (w/o enclosure) and conventional speaker

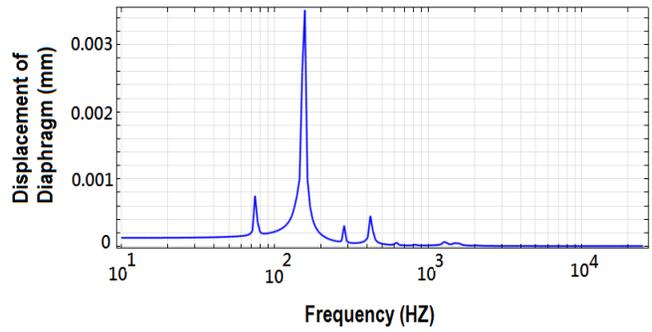


Fig. 10 Maximum displacement of FEPL diaphragm

Furthermore, the conventional speaker has its operational frequency within the low and middle frequency region while the FEPLs are better within the middle and high frequency region. In general, the conventional speaker (or FEPL) has the highest SPL in the low (or high) frequency region. Three additional cases of the FEPL as shown in Fig. 11 were investigated, i.e., the FEPL with enclosure and no vent (Case 1), the FEPL with enclosure and vented around the side walls (Case 2), and the FEPL with enclosure and vented on the diaphragm (Case 3). Fig. 12 shows the SPL performances of all the cases. Note that the FEPLs with enclosure showed a better performance than the infinite baffle one. However, for the FEPL with enclosure and no vent (Case 1), the performance is only effectual in the high frequency region, and the operational range is 1.3 to 20 kHz. The FEPLs with enclosure and vents around the side walls (Case 2) showed a better sensitivity extending towards the lower frequency region, the operational range is 60 Hz to 20 kHz and with a minimum average SPL of 50 dB in the lower frequency region and 90 dB in the higher frequency region. On the other hand, the performance Case 3 is between those of Cases 1 and 2.

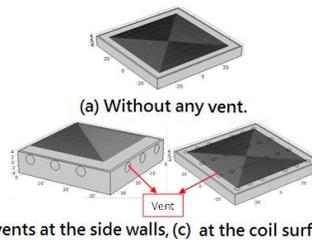


Fig. 11 FEPL with (a) no vent (b) vents around side walls, and (c) vents on diaphragm

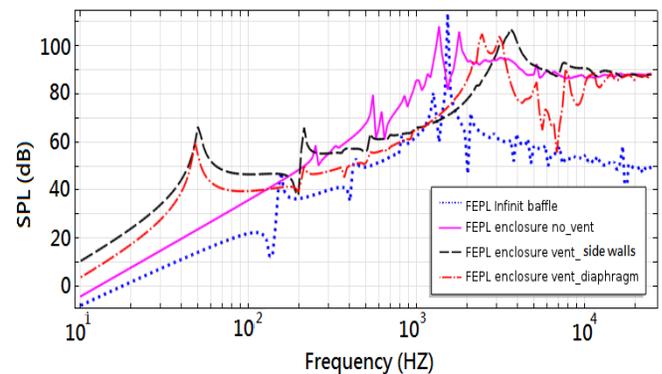


Fig. 12 SPLs for all the cases with and without enclosure and vents

## IV. CONCLUSION

This paper proposed a novel flexible electrodynamic planar loudspeaker that can generate a good audible sound. Some factors of the structure and configuration that led to tradeoffs are also made, such as the thickness of the cavity, magnet and coil, magnet configuration and polarization, and diaphragm dimension. The comparison of the FEPLs with the conventional cone type loudspeaker convincingly proved that the FEPLs have a better performance than the cone type loudspeaker in the high frequency region. In addition, the FEPL have other advantages over the cone type speaker, e.g. its flexibility, simple structure, cost effectiveness, easy manufacturability and its application versatility. The idea behind this study is a novel one, and as such has remained unexplored.

## ACKNOWLEDGMENT

This work was supported in part by National Science Council with the grants: NSC 101-2622-E-216-001-CC3, 101-2221-E-216-006-MY2, 101-2221-E-216-019-, and 102-2622-E-216-002-CC2.

## REFERENCES

- [1] M. Suzuki, T. Tsuzuki, T. Komiyama, T. Yamaguchi, T. Furukawa and S. Tokito, "Flexible colour OLED display based on phosphorescent material fabricated by ink-jet printing," in *Proc 13th IDW*, 2006, pp. OLED3-3–OLED3-6.
- [2] Y. Fujisaki, H. Sato, T. Yamamoto, H. Fujikake, S. Tokito and T. Kurita, "Flexible color LCD panel driven by low-voltage-operation organic TFT," *Journal of the Society for Information Display*, vol. 15, pp. 501–506, 2007.
- [3] S. Takehiro, O. Kazuho, A. Akio, K. Kohichi, H. Akira, M. Yuichi and M. Akito, "PVDF-driven flexible and transparent loudspeaker," *Applied Acoustics*, vol. 70, pp. 1021–1028, 2009.
- [4] R. Rashedin, T. Meydan and F. Borza, "Electromagnetic micro-actuator array for loudspeaker application," *Sensors and Actuators*, vol. 129, pp. 118–120, 2006.
- [5] C. S. Lee, J. Y. Kim, D. E. Lee, J. Joo, B. G. Wagh, S. Han, Y. W. Beag and S. K. Koh, "Flexible and transparent organic film speaker by using highly conducting PEDOT/PSS as electrode," *Synthetic Metals*, vol. 139, pp. 457–461, 2003.
- [6] C. H. Arved, B. Maxi, C. S. Georg, Z. Stefan, G. Andre and H. Christian, "Fully mass printed loudspeakers on paper," *Organic Electronics*, vol. 13, pp. 2290–2295, 2012.
- [7] J. M. Lin, "Eletro-acoustic transducer and method of manufacturing the same," U.S. Patent 2013/0163807 A1, June 27, 2013.
- [8] Fujifilm. Available: <http://www.diginfo.tv/v/13-0009-r-en.php>.
- [9] "Yamaha Develops Directional, Flat Panel Speakers - Video Inside," INAVATE, 7 April 2010. Available: <http://www.inavateonthenet.net>.
- [10] "SoundPad 580," Warwick audio technologies, Available: <http://www.warwickaudiotech.com/sites/default/files/downloads>.
- [11] C. Anthony, "ITRI paper-thin flexible loudspeaker won Wall Street Journal's Technology Innovation Awards," PRLog - Global Press Release Distribution, 06 June 2009. Available: <http://prlog.org/10365388>.
- [12] Z. Zhao, "Planar speaker system". US Patent 2013/0243238 A1, September 19, 2013.
- [13] S. S. Je and C. Junseok, "An electromagnetically actuated micromachined loudspeaker for hearing aids applications," in *Sensors, 2007 IEEE*, Atlanta, GA, 2007.
- [14] D. Graebener, "Single end planar magnetic speaker," U.S. Patent 7251342 B2, July 31, 2007.
- [15] G. Lemarquand, R. Ravaud, I. Shahosseini, V. Lemarquand, J. Moulin and E. Lefeuvre, "MEMS electrodynamic loudspeakers for mobile phones," *Applied Acoustics*, vol. 73, pp. 379–385, 2012.
- [16] I. Shahosseini, E. Lefeuvre, J. Moulin, E. Martincic, M. Woytasik, G. Pillonnet and G. Lemarquand, "Planar microcoil optimization of MEMS electrodynamic microspeakers," *IEEE Trans. Magnetics*, vol. 49, pp. 4843–4850, 2013.
- [17] C. H. Liou, M. D. Chen, "Flexible speaker," U.S. Patent 20090060249 A1, March 5, 2009.
- [18] F. Kontomichos, A. Koutsioubas, J. Mourjopoulos, N. Spiliopoulos and A. Vradis, "A thermoacoustic device for sound reproduction," in *Acoustics '08 Paris*, Paris, 2008.
- [19] L. Xiao, Z. Chen, C. Feng, L. Liang, Z. Q. Bai, Y. Wang, L. Qian, Y. Zhang, Q. Li, K. Jiang and S. Fan, "Flexible, stretchable, transparent carbon nanotube thin film loudspeakers," *Nano Letters*, vol. 8, pp. 4539–4545, 2008.
- [20] D. Mat, "electronicdesign," 13 December 2009. Available: <http://electronicdesign.com/boards/thin-speaker-technology>.
- [21] Arnold Magnetic Technologies Corp., "FLEXMAG," Arnold Magnetic Technologies, 2014. Available: <http://www.arnoldmagnetics.com>.
- [22] L. L. Beranek and T. J. Mellow, "Electrodynamic loudspeakers," in *Acoustics: Sound Fields and Transducers*, Academic Press, 2012, pp. 241–288.
- [23] "Loudspeaker Driver," COMSOL Multiphysics, 2013. Available: [https://www.comsol.com/model/download/39132/loudspeaker\\_driver](https://www.comsol.com/model/download/39132/loudspeaker_driver)
- [24] R. R. Daniel, *The Science and Applications of Acoustics*, 2nd ed., Springer Science Business Media Inc., 2006, p. 583.

**Jium-Ming Lin** was born at Taipei, Taiwan in 1952. Prof. Lin was graduated from the department of Electronic Engineering, National Chiao-Tung University at Hsin-Chu, Taiwan in 1974. He had also achieved the Master and Ph. D Degrees from the same school of Institute of Electronics in 1976 and 1985, respectively. Dr. Lin was a researcher at Chung-Shan Institute of Science and Technology in Taiwan from 1978 to 1992. His major field was in surface-to-air missile navigation, guidance and control. He has been an adjunct professor and full professor since 1992 and 1996 at Dept. of Mechanical Engineering, Chung-Hua University, Taiwan. He was the director of Dept. of Mechanical Engineering from 1996 to 1997; Prof. Lin He was also the director of R&D of Chung-Hua University. He has been as a Professor at Dept. of Communication Engineering from 2009. Prof. Lin majors in the fields of RFID, wireless accelerometer and angular accelerometer, multi-variable control, optimal control, stochastic control, fuzzy control, avionics, MEMS, semiconductor fabrication and packaging, measurement and mechatronics.

**Cheng-Hung, Lin** was born at Taipei, Taiwan in 1985. Mr. Lin was graduated from the Department of Mechanical Engineering, Chung-Hua University at Hsin-Chu, Taiwan in 2009. He had also achieved the Master Degree from the same school in 2012. His major field was in missile navigation, guidance and control. The other interests are as RFID, wireless accelerometer, angular accelerometer, multi-variable control, optimal control, stochastic control, fuzzy-neural control, avionics, and MEMS design.

# Polarization Angle Independent Perfect Metamaterial Absorber

C. Sabah, F. Dincer, E. Demirel, M. Karaaslan, E. Unal and O. Akgol

**Abstract**— This paper presents a perfect metamaterial absorber (MA) based on rings and cross wires (RCWs) configuration in microwave frequency regime. Maxima absorption rate is 99.9% at 2.76 GHz for simulation and 99.4% at 2.82 GHz for experiment, in order. The proposed MA provides perfect absorption with angle of polarization independency. Consequently, suggested model enable myriad potential applications such as stealth and military technologies.

**Keywords**— perfect absorber; metamaterial; microwave.

## I. INTRODUCTION

MTMs have unconventional electromagnetic (EM) properties, such as artificial magnetism and negative refraction. They still draw interest of scientists due to practical importance owing to varieties of potential application areas [1]. These materials are manmade and can be artificially fabricated at the desired regimes of the EM spectrum from MHz to near-IR [2-7]. Also, MTMs have many interesting applications which are not found in conventional materials, such as super lens, sensing, cloaking, waveguide, absorber and so on.

In this paper, we evaluated previous MA studies in the literature. Then, we designed a new perfect MA that shows polarization angle independency with perfect absorptivity in the microwave frequency. Besides, presented advantages of the suggested MA model are evaluated in detailed.

## II. THEORETICAL APPROACH

The absorption value with respect to frequency is defined as  $A(\omega) = 1 - R(\omega) - T(\omega)$ , where  $A(\omega), R(\omega) = |S_{11}|^2$  and  $T(\omega) = |S_{21}|^2$  represent the absorption, reflection and transmission of the system, in order. The reflection and transmission values decay at a desired frequency range in an

M.K. acknowledges the support of TUBITAK under the Project Number of 113E290 and partial support of the Turkish Academy of Sciences

C. Sabah is with the Department of Electrical and Electronics Engineering, Middle East Technical University - Northern Cyprus Campus, Kalkanli, Guzelyurt, TRNC / Mersin 10, Turkey (e-mail: sabah@metu.edu.tr).

F. Dincer is with the Department of Computer Engineering, Mustafa Kemal University, Iskenderun, Hatay, 31200, Turkey

E. Demirel is with TUBITAK - UME, 41470 Gebze, Kocaeli, Turkey.

M. Karaaslan, E. Unal and O. Akgol are with the Department of Electrical and Electronics Engineering, Mustafa Kemal University, Iskenderun, Hatay, 31200, Turkey.

absorber due to the impedance matching and metallic background surface. In the maximum absorption condition, the effective impedance of the overall system ( $Z(\omega) = \sqrt{\mu(\omega)/\varepsilon(\omega)} = z_1 + iz_2$ ) exactly matches with the free space impedance  $Z(\omega) = Z_0(\omega)$  and the reflection decays [8].

## III. NUMERICAL STUDY, RESULTS, AND DISCUSSION

The designed MA consists of RCWs-shaped, dielectric and metallic layer. The metallic structures on the top and bottom layers of the substrate are modeled as copper sheet with electrical conductivity of  $5.8 \times 10^7$  S/m and thickness of 0.035 mm. The thickness, loss tangent, relative permittivity and permeability of the selected dielectric-FR4 are 1.6 mm, 0.02, 4.2 and 1, respectively. The top resonator, bottom metallic plate and FR4-substrate constitute the MA. The dimensions of the RCWs-shaped inclusion are introduced as shown in Fig. 1(a). After simulations (by CST Microwave Studio based on finite integration technique), the MA is manufactured as shown in Fig. 1(b). Also, the measurement of S-parameters is achieved by using ROHDE & SCHWARZ ZVL6 VNA. The VNA supplies microwaves in the range of 1 GHz - 6 GHz through two horn antennas as shown in Fig. 1(c) is the schematic view of the measurement setup. Note that the transmitter antenna and the front side of the sample are arranged to form face-to-face configuration with each other in the measurement.

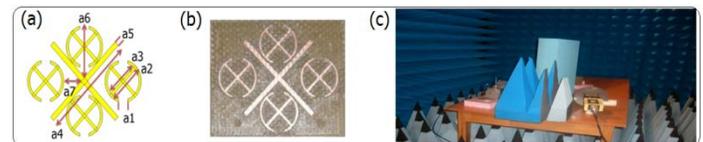


Fig. 1 (a) Represent dimensions of the structure ( $a_1=2$  mm,  $a_2=8$  mm,  $a_3=7$  mm,  $a_4=20$  mm,  $a_5=1$  mm,  $a_6=10$  mm,  $a_7=3.5$  mm), (b) picture of the manufactured microwave scale sample, (c) measurement setup of the system

Numerical and experimental results are proved that the suggested model is very well candidate for perfect MAs as shown in Fig. 2. Maximum absorption rate is observed approximately 99.9% at 2.76 GHz for simulation and 99.4% at 2.82 GHz for experiment, respectively. Also, fractional bandwidth (FBW) calculations of the resonance region are examined to show the qualification of the proposed MA. Since, bandwidth calculations are crucially important for

numerous applications. It is obtained by the formula  $FBW = \Delta f / f_0$ , where  $\Delta f$  and  $f_0$  represent the half power bandwidth and the center frequency, respectively. In the suggested MA, these are obtained as  $\Delta f = 0.122$  GHz,  $f_0 = 2.76$  GHz,  $FBW \approx 4.42\%$  for simulation and  $\Delta f = 0.1$  GHz,  $f_0 = 2.82$  GHz,  $FBW \approx 3.54\%$  for experiment. Moreover, the proposed structure has approximately 120 MHz bandwidth range referring to 4.42% FBW which is quite enough for many applications. For example if we want to use a patch antenna which has approximately 3% FBW in an application, our structure would provide enough margins to work with. These calculations show performance quality of the suggested MA.

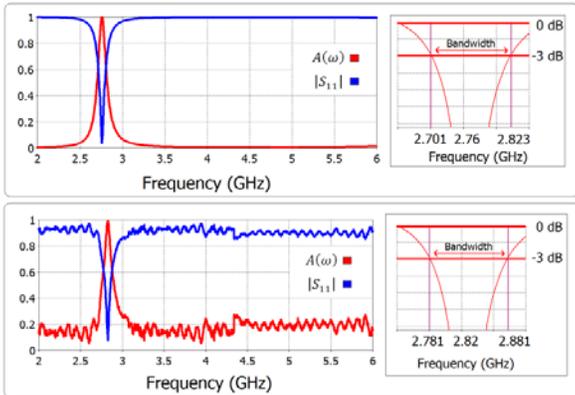


Fig. 2 Simulated (left-side) and measured (right-side) reflection & absorption and the FBW for the proposed MA system.

As the next investigation, we analyzed the effects of different polarization angles for the proposed MA. For this reason, the RCW-shaped inclusion is numerically rotated from to  $90^\circ$  with  $15^\circ$  steps as shown in Fig. 3. It can be seen that the suggested MA model provides very good absorption for all polarization angles due to the structural symmetry. In addition, when the polarization angle is changed, the MA provides dual band absorption between the frequency ranges of 2.5-3 GHz and 4.5-5 GHz. Shifts in the resonance frequencies are still very small with respect to the normal incidence case (Fig. 2). Also, the additional peaks increased or decreased depending on polarization angles.

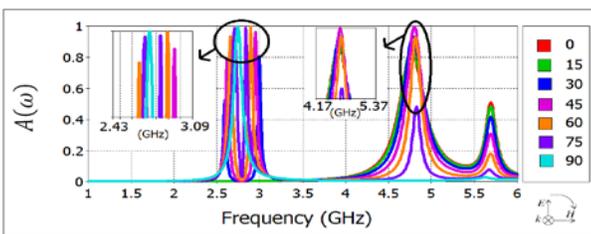


Fig. 3 Simulation absorption characteristics at different polarization angles from  $0^\circ$  to  $90^\circ$ .

In order to view the physical mechanism of the operation principle of the proposed model at the resonance frequency, the electric field and surface current distributions of the MA are investigated as shown in Fig. 4, respectively. High electric field concentration is happen around upper and lower rings,

cross wires and gap between them verify absorber resonance mode. The electric field causes to excitation of surface charge throughout the same path. Hence magnetic dipole moment due to surface charge induces magnetic response and leads to resonance absorption. E and H components of the incident EM wave are generated by strongly couples of these responses at the resonance frequency. This provides both electric and magnetic resonance at resonance frequency.

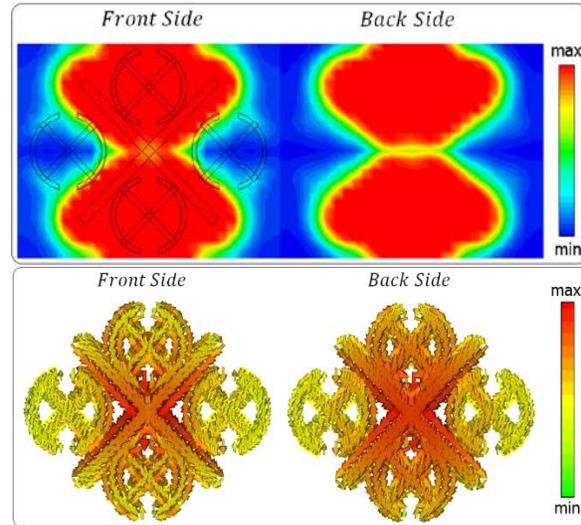


Fig. 4 Electric field and surface current distribution at resonance frequency of 2.76 GHz.

#### IV. CONCLUSION

The study presents design, simulation and experiment a perfect MA. We designed, simulated and measured a perfect MA. Obtained results are discussed and compared. Simulation results are in good agreement with the experiment results. Also it can be seen from obtained results, this structure can be used as both angle and polarization independent absorber. By scaling the dimensions, the proposed model in microwave can be applied to other frequency regime. Moreover, the suggested model can be used in short wave communication and pressure measurement applications.

#### REFERENCES

- [1] F. Dincer, C. Sabah, M. Karaaslan, E. Unal, M. Bakir, and U. Erdiven, "Asymmetric transmission of linearly polarized waves and dynamically wave rotation using chiral metamaterial," *Prog. Electromagn. Res.*, vol. 140, pp. 227, 2013.
- [2] M.C.K. Wiltshire, J.B. Pendry, I.R. Young, D.J. Larkman, D.J. Gilderdale, and J.V. Hajnal, "Microstructured Magnetic Materials for RF Flux Guides in Magnetic Resonance Imaging," *Science*, vol.291, pp. 849, 2001.
- [3] D.R. Smith, W.J. Padilla, D.C. Vier, S.C. Nemat-Nasser, and S. Schultz, "Composite Medium with Simultaneously Negative Permeability and Permittivity", *Phys. Rev. Lett.*, vol. 84, pp.4184, 2000.
- [4] M. Gokkavas, K. Guven, I. Bulu, K. Aydin, R.S. Penciu, M. Kafesaki, C.M. Soukoulis, and E. Ozbay, "Experimental demonstration of a left-handed metamaterial operating at 100 GHz," *Phys. Rev. B*, vol. 73, pp. 193103, 2006.

- [5] T.J. Yen, W.J. Padilla, N. Fang, D.C. Vier, D.R. Smith, J. B. Pendry, D. N. Basov, and X. Zhang, "Terahertz Magnetic Response from Artificial Materials," *Science*, vol. 303, pp.1494, 2004.
- [6] Linden, C. Enkrich, M. Wegener, J. Zhou, T. Koschny, and C. M. Soukoulis, "Magnetic response of metamaterials at 100 Terahertz", *Science*, vol. 306, pp.1351, 2004.
- [7] S. Zhang, W. Fan, N.C. Panoiu, K. J. Malloy, R.M. Osgood, and S.R.J. Brueck, "Experimental demonstration of near-infrared negative-index metamaterials," *Phys. Rev. Lett*, vol. 95, pp. 137404, 2005.
- [8] F. Dincer, M. Karaaslan, E. Unal, and C. Sabah, "Dual-band polarization independent metamaterial absorber based on omega resoanator and octa-star strip configuration," *Prog. Electromagn. Res*, vol. 141, pp. 219, 2013.

# On Adaptation Possibility of Model Based on Slow Flow Around Sphere for Determination of Flow Local Speeds in Window Between Spheres

SANDULYAK ANNA, SANDULYAK DARYA, SEMINA OLGA, SANDULYAK ALEXANDER  
 Moscow State University of Instrument Engineering and Computer Science (MGUPI)  
 20 Stromynka, Moscow, 107996  
 RUSSIAN FEDERATION  
 anna.sandulyak@mail.ru <http://www.mgupi.ru>

**Abstract:** - Considered here is a version of Newtonian liquid flow model in window between spheres of granular medium based on classical model of slow flow around isolated sphere adapted for such problem. With justified ignoring flow speed radial component, and qualitative and quantitative modification of speed parameter of flow running on sphere, the corresponding correction factor and expressions are defined for flow local speed at any point of window. Given here are considerably larger Reynolds's numbers limiting applicability of expressions obtained (in comparison with basic ones).

**Key-Words:** - local speed, average speed, filtering speed.

## 1 Introduction

The problem of studying flow speed profile in granular medium, i.e. in specific pores-channels of variable section, fairly is considered difficult even in the event of special case, for example, in slow (creeping) flow of incompressible liquid (when inertia members in Navier-Stokes's equation are negligible) in characteristic "star-shaped" window section (Fig. 1) between contacting granules-spheres. Thus,

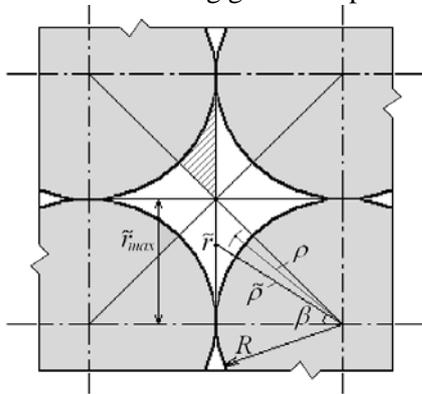


Fig.1. Contacting spheres with window between them, flow normal to window.

model combination of granules-spheres shown in Fig.1 forming considered window despite simplification adequately is quite accepted: It is inherent not only to classical packing of granules-spheres with cubic structure, but also with more dense packing. So, in polyspherical medium structure with "fractional" square-rhombic cells [1] (this structure as regards packing density and related porosity, and also as regards coordination number is more close to natural packing-filling) four "faces" from six – squares and only two – rhombs.

In our opinion, the approach based on classical model of slow flow around isolated sphere could be one of approaches to solution of the task set here.

At first sight, this approach possibility can seem to be inappropriate in view of obvious striking difference in conditions of flow around isolated sphere and sphere in sphere group.

Meanwhile, using basic provisions of classical model as representing fundamental nature of slow flow behavior in the vicinity of spherical surface seems to be quite justified here, at least – near spheres at a distance not exceeding half the value of interspherical gaps. Thus, realization of the approach demands, naturally, corresponding reasonable estimates and assumptions.

As for those concrete basic provisions of classical model, which could be useful in solution of the problem considered, first, these are solutions concerning tangential  $u_\theta$  and radial  $u_\rho$  speed components of liquid flowing round sphere. In spherical coordinate system the expressions for these, as we know, are as follows:

$$\begin{aligned} u_\theta &= -v \cdot \sin \theta \left( 1 - \frac{3R}{4\rho} - \frac{R^3}{4\rho^3} \right), \\ u_\rho &= v \cdot \cos \theta \left( 1 - \frac{3R}{2\rho} + \frac{R^3}{2\rho^3} \right), \end{aligned} \quad (1)$$

where:  $R$  – sphere radius,  $\rho$  – module of radius-vector emerging from sphere center of this or that point outside the sphere (Fig. 1),  $\theta$  – angle between selected flow direction and radius-vector of this point,  $v$  – speed of flow running on sphere (flow speed far from sphere, theoretically – at infinity).

## 2 Prerequisites to adaptation of classical expressions for speed in window between contacting spheres

### Comparison of tangential and radial speed components

In relation to considered model, the range of possible variation  $\rho/R$  is not traditionally wide (as for isolated sphere, when  $\rho/R=1\dots\infty$  or  $R/\rho=0\dots1$ ), but quite narrow limited to window sizes (Fig. 1):  $\rho/R=1\dots\sqrt{2}$  (i.e. up to window center) or similarly –  $R/\rho=0.71\dots1$ .

In similar variation range of  $\rho/R$ , the values of tangential  $u_\theta$  and radial  $u_\rho$  speed components are obviously inadequate. It is possible to be convinced in it preliminarily if to compare two expressions in (1):  $(1-3R/4\rho-R^3/4\rho^3)$  and  $(1-3R/2\rho+R^3/2\rho^3)$ , from which the first considerably prevails (Fig. 2). As for  $u_\theta$  and  $u_\rho$  values, more exact – their modules  $|u_\theta|$  and  $|u_\rho|$ , at  $\theta=45^\circ$  and  $\theta=135^\circ$ , when  $|\sin\theta|=|\cos\theta|$ , distinction between  $|u_\theta|$  and  $|u_\rho|$  corresponds to mutual distinction of mentioned expressions (Fig. 2). Moreover, at  $45^\circ<\theta<135^\circ$ , when  $|\cos\theta|<|\sin\theta|$ , distinction between  $|u_\theta|$  and  $|u_\rho|$  becomes still more; in the window plane (Fig. 1), when  $\cos\theta=0$ , in general  $|u_\rho|=0$ .

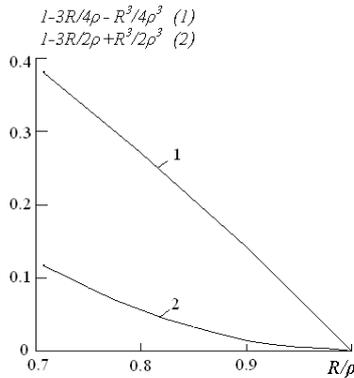


Fig.2. Comparison of expressions in brackets of equations (1) for characteristic (in relation to considered problem) range  $R/\rho$ : from  $R/\rho=0.71$  (at a distance from sphere surface, equivalent to distance to window center between spheres) to  $R/\rho=1$  (on sphere surface).

Thus, in the range  $45^\circ\leq\theta\leq135^\circ$ , i.e. and in the window plane, and in quite large areas at both sides of this plane, the module of tangential speed  $|u_\theta|$  is obviously dominating. Thus, with an error not exceeding 5% (and that – outside the window plane: at  $\theta=45^\circ$  and  $\theta=135^\circ$ , when  $|\sin\theta|=|\cos\theta|$ ), the resulting speed module calculated in general case as

$u = (u_\rho^2 + u_\theta^2)^{0.5}$ , practically corresponds to value of  $|u_\theta|$ :

$$u \cong |u_\theta| = v \cdot \sin\theta \left( 1 - \frac{3R}{4\rho} - \frac{R^2}{4\rho^3} \right),$$

$$u \cong 1.3v \cdot \sin\theta \left( 1 - \frac{R}{\rho} \right). \quad (2)$$

In addition, the second version of speed equation is given here, which is simpler and almost equivalent (in the range specified for window  $R/\rho=0.71\dots1$ ). For this purpose, the expression in brackets of the basic (first) equation (2) is artificially replaced with corresponding linear function (Fig. 2):  $1-3R/4\rho-R^3/4\rho^3 \cong 1.3(1-R/\rho)$ .

From (2), as special cases there also follow equations of flow local speed at any point in the window plane ( $\sin\theta=1$ ):

$$u = v \left( 1 - \frac{3R}{4\rho} - \frac{R^3}{4\rho^3} \right), \quad u \cong 1.3v \left( 1 - \frac{R}{\rho} \right). \quad (3)$$

### Speed of flow running on sphere – as correction parameter

Certainly, for polyspherical medium (Fig. 1) the flow speed running on sphere  $v$  used in (1)-(3) completely loses its initial meaning (we will notice: never "transforming" into the speed of flow running on polyspherical medium, i.e. into the so-called filtering speed  $v_f$ ).

But parameter  $v$ , being purely spurious here (according to name), should become quite certain numerical parameter with accuracy true for these equations. Thus, parameter  $v$  as an average correction factor, naturally should be definitely coordinated with flow speed characteristics in polyspherical medium, e.g. with average flow speed in window  $\langle u \rangle$ , filtering speed  $v_f$ .

To substantiate parameter  $v$ , in particular, for definition of key relationship between  $v$  and  $\langle u \rangle$ , it is sufficient, having limited to one of eight identical window "sectors" (one of them is hatched in Fig. 1), in its limits to perform corresponding integration averaging of local speed  $u$  expressed by the first (main) equation (3). Thus, it is evident that double integration is required in this case: at first by  $\rho$  from  $R$  to  $\tilde{\rho}$ , and then by  $\beta$  from 0 to  $\pi/4$  (Fig. 1), i.e.

$$\langle u \rangle = \frac{1}{\pi/4} \int_0^{\pi/4} d\beta \cdot \frac{1}{\tilde{\rho} - R} \int_R^{\tilde{\rho}} u \cdot d\rho, \quad (4)$$

taking into consideration (Fig. 1) that  $\tilde{\rho} = R/\cos\beta$  – radius-vector module of any point located on window diagonal, and  $\beta$  – angle between its radius-vector and intercenter line of spheres.

The first integration (4) leads to this intermediate result:

$$\langle u \rangle = \frac{4v}{\pi} \int_0^{\pi/4} \left( 1 + \frac{3 \cos \beta \cdot \ln \cos \beta}{4(1 - \cos \beta)} - \frac{1}{8} \cos^2 \beta \right) d\beta \quad (5)$$

Further integration also does not cause difficulties except that a lengthy function  $\cos \beta \cdot \ln \cos \beta / (1 - \cos \beta)$  in the second member of the integrand can be integrated only using methods of approximate calculation (e.g., Simpson's formula). At the same time, it is possible to proceed differently: this function absolutely without accuracy reduction (Fig. 3) is replaced with equivalent function  $(0.25\beta^2 - 1)$  more convenient for integration, of course, in the necessary interval  $\beta=0 \dots \pi/4$  ( $\beta=0 \dots 0.785$ ).

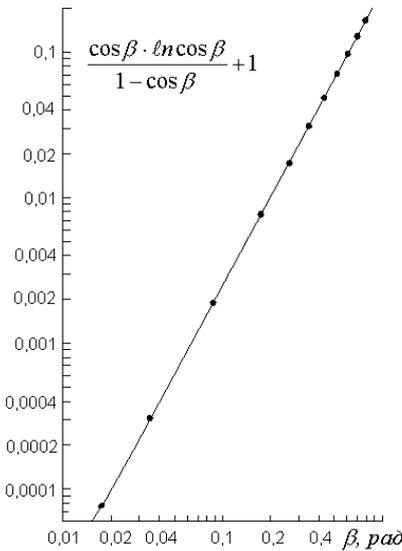


Fig.3. Illustration of possibility (in logarithmic coordinates) of almost equivalent replacing expression  $\cos \beta \cdot \ln \cos \beta / (1 - \cos \beta) + 1$  (shown as dots) with exponential function  $0.25\beta^2$  (shown as line).

Then integration (5) leads to final result:

$$\langle u \rangle = \frac{4v}{\pi} \int_0^{\pi/4} \left[ 1 + \frac{3}{4} (0.25\beta^2 - 1) - \frac{1}{8} \cos^2 \beta \right] \cdot d\beta$$

$$= \frac{v}{13.5} \quad (6)$$

Therefore, for flow model in window (Fig. 1) this parameter "introduced" from classical model as speed of flow running on sphere  $v$  gets the following quantitative interpretation:

$$v = 13.5 \cdot \langle u \rangle \quad (7)$$

### 3 Form of expressions adapted (to window between spheres) for flow local speed

Taking into account modified parameter  $v$  of expression (3) defined in (7) for local  $u$  (referred to average  $\langle u \rangle$ ), the flow speed in window can be represented as follows:

$$\frac{u}{\langle u \rangle} = 13.5 \left( 1 - \frac{3R}{4\rho} - \frac{R^3}{4\rho^3} \right) \cong 17.6 \left( 1 - \frac{R}{\rho} \right), \quad (8)$$

with maximum value of flow speed  $u = u_{max}$  in window, i.e. in its center with coordinate  $\rho/R = (\tilde{\rho}/R)_{max} = \sqrt{2}$  (Fig.1):

$$u_{max} = 5.1 \cdot \langle u \rangle \quad (9)$$

Thus, equations (8) can be also represented as follows:

$$\frac{u}{u_{max}} = 2.64 \left( 1 - \frac{3R}{4\rho} - \frac{R^3}{4\rho^3} \right) \cong 3.4 \left( 1 - \frac{R}{\rho} \right), \quad (10)$$

where  $u/u_{max}$  values variate from 0 to 1 at variation of  $\rho/R$  from 1 to  $\sqrt{2}$ .

Included in (8) and (10), parameters  $\langle u \rangle$  and  $u_{max}$  interconnected via (9) can be easily expressed, in particular, through filtering speed  $v_f$  (speed of flow running on porous medium). So, as for relationship of  $\langle u \rangle$  and  $v_f$ , it will be shown as

$$\langle u \rangle = 4.7 \cdot v_f \quad (11)$$

when accepting identity condition of "flow capacity" in formal and actual window (Fig. 1):  $v_f \cdot d^2 = \langle u \rangle \cdot (d^2 - \pi d^2/4)$ .

Then, equation (8) gets the form:

$$\frac{u}{v_f} = 63 \left( 1 - \frac{3R}{4\rho} - \frac{R^3}{4\rho^3} \right) \cong 82 \left( 1 - \frac{R}{\rho} \right), \quad (12)$$

rather convenient for practical calculations.

### 4 Summary

Expression (12) can be used in solution of a wide range of problems, for example, in theoretical and technological feasibility of magnetophoresis filtration process (practiced for effective extraction of ferroparticles [2-6]) for required profound analysis not only of magnetic, but also of a number of competing forces, in particular, Stokes force. Here, information on flow speed profile becomes very useful.

Besides, the additional condition concerning assessment of Reynolds limit number limiting application of solutions obtained above is essential.

Initial (for development of the model considered here) classical expressions (1) are certainly true for values of Reynolds numbers up to  $Re = Re(v, d) = 1-2$  calculated for isolated sphere always

explicitly: on the flow speed  $v$  running on sphere and sphere diameter  $d$ .

However, it is possible to assume that this value of the number (limiting for flow around isolated sphere – when resistance coefficient is still inversely proportional to  $Re$ ) –cannot be even used as reference point of Reynolds limit number for flow in polyspherical medium pores. Thus, though this number  $Re$  is calculated, apparently, using the same (but formal for porous medium, i.e. in relation to pores-channels of this medium) parameters, namely on filtering speed  $v_f$  (speed flow of running on porous medium) and sphere diameter  $d$ , it is really necessary to insist on this statement.

So, for liquid flow in granular medium the limit number  $Re = Re(v_f, d)$ , limiting applicability of Darcy's law (before noticeable manifestation of inertia effects), is often estimated by considerably higher values. For example, it can be estimated on dependence of pressure losses in this medium from number  $Re(v_f, d)$ . It is characteristic that it remains linear up to values  $Re(v_f, d) = 60-80$  [7] showing thereby a "prolonged" retaining of laminar flow mode (thus, resistance coefficient in the known Darcy-Weisbach formula for pressure losses remains inversely proportional to this number, i.e. inertia effects are manifested not so considerably).

## 5 Acknowledgements

This work was supported by Russian Federation Ministry of Education and Science.

### References:

- [1] Sandulyak A.V., Sandulyak A.A., Yershova V.A. Functional correction to classical expression for average flow speed in granular densely packed medium, *Theoretical bases of chemical technology*, 42, 2008, No 2, pp. 231-235.
- [2] Araj S., Moyer C.A., Aidun R., Matijevic E. Magnetic filtration of submicroscopic particles through a packed bed of spheres, *Journal of Applied Physics*, 57, 1985, pp.4286.
- [3] Watson J., Watson S. The ball matrix magnetic separator, *IEEE Transactions on Magnetics*, V.19, Issue 6, 1983, pp.2698-2704.
- [4] Zezulka V., Straka P., Mucha P. A magnetic filter with permanent magnets on the basis of rare earth, *Journal of Magnetism and Magnetic Materials*, 268, 2004, pp. 219-226.
- [5] Sandulyak A.A., Sandulyak A.V. Application prospects of magnetic filters-separators for purification of ceramic suspensions, *Glass and ceramics*, 11, 2006, pp. 34-37.
- [6] Svoboda J. A realistic description of the process of high-gradient magnetic separation, *Minerals Engineering*, 14, 2001, pp.1493-1503.
- [7] Sandulyak A.V., Plaul P., Marr R., Gamze T. Exponential nature of pressure losses in granular media, *Heavy mechanical engineering*, 6, 2002, pp. 20-25.

# Neural Networks Based Feature Selection from KDD Intrusion Detection Dataset

Adel Ammar, Khaled Al-Shalfan  
 College of Computer Sciences and Information  
 Computer Sciences Department  
 Al-Imam Mohammad Ibn Saud Islamic University  
 Riyadh, KSA  
 adel.ammar@ccis.imamu.edu.sa

**Abstract**—We present the application of a distinctive feature selection method based on neural networks to the problem of intrusion detection, in order to determine the most relevant network features. We use the same procedure for feature selection and for attack detection, which gives more consistency to the method. We apply this method to a case study and show its advantages compared to some existing feature selection approaches. We then measure its dependence to the network architecture and the learning database.

**Keywords**—Intrusion detection, network security, feature selection, KDD dataset, neural networks.

## I. INTRODUCTION

For Intrusion Detection Systems (IDS), ranking the importance of input features is a problem of significant interest, since the elimination of irrelevant or useless inputs leads to a simplification of the problem and may allow faster and more accurate detection. This is especially critical for the construction of an efficient real-time IDS able to comply with the constraints of high speed networks. We present, in this article, a feature selection method based on Neural Networks (NN), classifying traffic features according to their relative contribution to attack detection.

Section II introduces the method and describes its theoretical basis. Section III details the results of a case study for a single output classification NN, and reviews the advantages and limitations of the method. Finally, section IV draws a conclusion for

the present work and mentions some open issues for future works.

## II. THEORETICAL BASIS

The method we propose here for selecting connection features is based on feed-forward neural networks. It has been applied in another application by [1] and theoretically formulated by [2] who called it HVS (Heuristic for Variable Selection). Nevertheless, it has not yet been applied to intrusion detection, to the best of our knowledge.

We introduce the features that need to be ranked as inputs of a feed-forward neural network (with a single hidden layer) used as a classifier that distinguishes attacks from normal traffic. After the training process on a representative learning database, we assess the relative contribution of each feature as follows. We expect the contribution  $C_{js}$  of a neuron  $j$  of the hidden layer to the output  $s$  according to the formula:

$$C_{js} = \frac{|W_{js}|}{\sum_{k=1}^{N_h} |W_{ks}|} \quad (1)$$

Where  $W_{ks}$  is the weight of the connection between a hidden neuron  $k$  and the output  $s$  and  $N_h$  is the number of hidden neurons. Then, we obtain the contribution of an input neuron  $i$  to the output according to the formula:

$$C_{is} = \sum_{j=1}^{N_h} C_{js} \cdot \frac{|W_{ij}|}{\sum_{k=i}^{N_i} |W_{kj}|} \quad (2)$$

Where  $W_{ij}$  is the weight of the connection between the input neuron  $i$  and a hidden neuron  $j$  and  $N_i$  is the number of inputs. The sum of input contributions is, therefore, equal to 1.

### III. CASE STUDY ON KDD DATABASE

#### A. Calculation of features' contribution

We have applied the HVS method described above, in a case study, to the KDD 99 intrusion detection benchmark [3]. This database originated from the 1998 DARPA Intrusion Detection Evaluation Program that was prepared and managed by MIT Lincoln Labs. The objective was to assess and evaluate research in intrusion detection [4]. The dataset was summarized into network connections with 41 features per connection. In order to measure the relevance of these features, we constructed a NN with a single output that distinguishes between normal traffic and attacks. The learning database used to train the NN consists of a 1% random extraction (4,940 samples) from the original KDD learning set (containing 494,021 connection records). A learning database with such a size is sufficient to achieve an accuracy rate of 92% on the KDD test set (composed of 311029 independent connection records).

Figure 1 depicts the obtained results, after applying the HVS method following (1) and (2). Features # 20 and 21 take a null contribution because they are constant in the whole KDD learning set. The same can be noticed for features # 9 and 15, which are almost constant. In fact, more than 99.999% of the KDD learning set connection records contain a null value for these two features. Features 7, 11 and 18 could also be excluded from the learning database since their contribution is remarkably little; while the most significant features are # 10, 22, 23, 34, 36, 39.

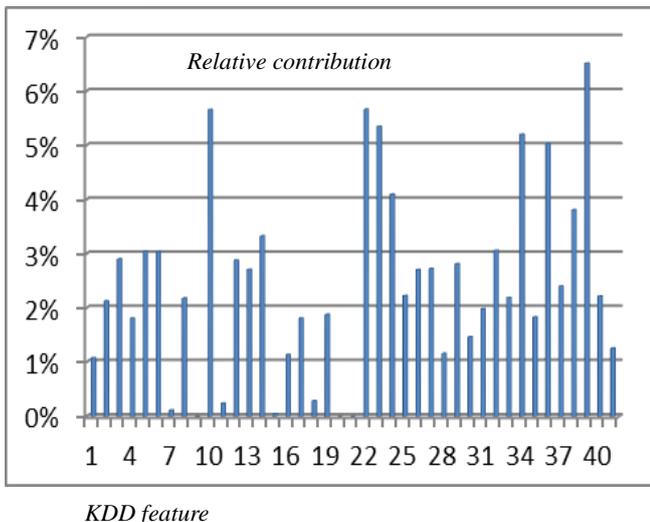


Figure 1. Relative contribution of each of the KDD 41 features to the detection of attacks (distinction between normal traffic and attacks of various types)

#### B. Checking the consistence of the method

In order to verify the consistence of the results, we selected a set of most significant features (calculated as in the section above) to be set as inputs of the classification NN, and compared the results with those obtained with the full set of inputs. Figure 2 shows these results after applying the networks to the testing databases. We note that we can keep only the most influential 12 features (out of 41), without significantly deteriorating neither the overall accuracy rate (Figure 2) nor the false positive and false negative rates (Figure 3).

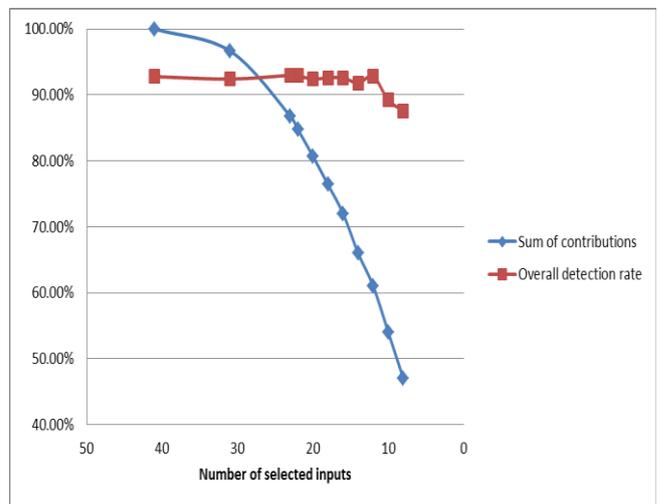


Figure 2. Evolution of the overall accuracy rate according to the number of selected inputs

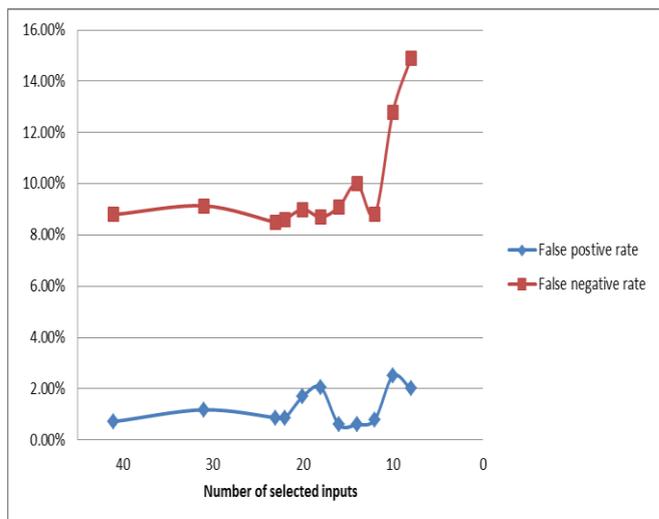


Figure 3. Evolution of the false positive rate according to the number of selected inputs

c. *Advantages of the method*

The results shown above are consistent with those obtained by [5] and [6]. The latter used a totally different method which consists in deleting one of the features and measuring its impact on the result, using either a Neural Network or an SVM classifier. Compared to this approach, the method we have presented above shows several advantages:

- The deletion-based method needs to run as many trainings as the number of features, each time deleting one of the features while the HVS method ranks all the features after a unique training, and does not imply any complicated computation.
- The HVS method tends to be more accurate in selecting relevant features than the method used by [6] as explained in section III.A.2.
- The HVS method distinguishes well between features than the SVM based feature ranking used by [6] which yields remarkably close accuracy results for most of the features, with so slight variations that they could be of random origin.
- The HVS method reveals to be more precise in detecting irrelevant features than the method presented in [6]. For example, while features 20 and 21 are constant in the whole KDD learning dataset (as previously noticed by [5]), and features 9 and 15 almost constant and they were not detected as the least important features in [6].

On the other hand, in term of consistence of HVS method, we note that we can keep only the most important 12 features (out of 41), without significantly deteriorating neither the overall accuracy rate (Figure 2) nor the false positive and false negative rates (Figure 3). This number of features is close to the one retained by [9] (11 features) using rough sets and genetic algorithms. [6] conducted a similar test but showed a significant deterioration when selecting the most important 34 features (the overall accuracy rate decreased from 87% to 81% and the false positive rate increased from 6.7% to 18%). This tends to prove that our selection feature method is considerably more accurate than other cited methods. It should be also noticed that these latter results shown by [6] are not consistent with the Figures they obtained during the feature ranking since the deletion of only one feature (#10 or #35) decreased the accuracy of their network to less than 55%. They did not precise on which database they tested their result. intuitively the results they gave for the SVM classification suggests that they tested on only a part of the KDD training dataset (so with a very close distribution to that of the learning database) while we tested on the

independent KDD testing dataset (which an entirely different distribution of attacks, and containing new attack types), which is more realistic. Obviously, testing on the training data set yields an artificially high performance.

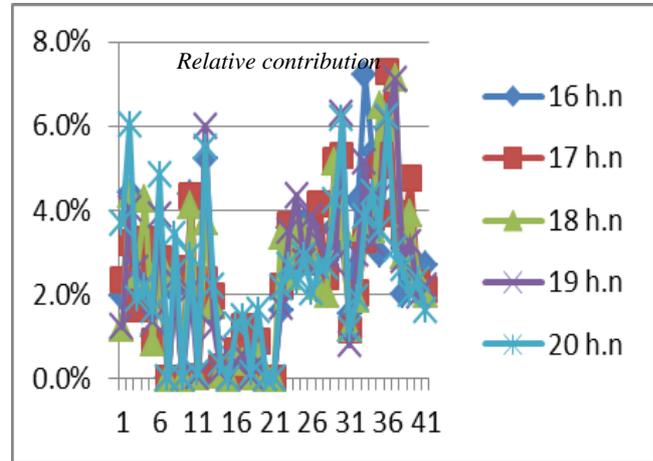


Figure 9. Relative contribution of each of the KDD 41 features to the detection of normal traffic, calculated for five different networks (with a number of hidden neuron varying from 16 to 20)

Furthermore, the contributions of the inputs, calculated using the HVS method, are largely independent of the network architecture, as shown in Figure 9. This Figure depicts the result of use of the HVS method to five networks with different internal architectures. The five tests show very close results. Nevertheless, this stands only if the number of hidden neurons is sufficient to resolve the classification problem.

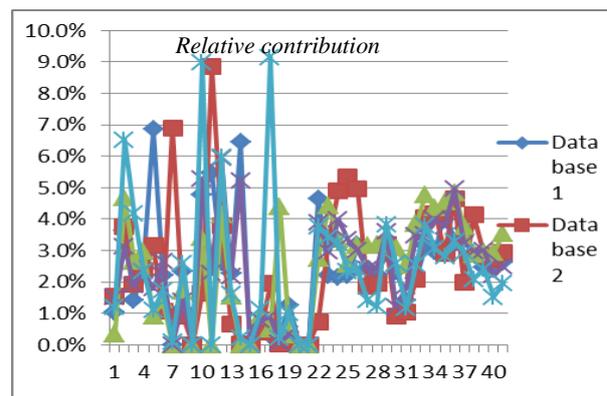


Figure 10. Relative contribution of each of the KDD 41 features to the detection of normal traffic, calculated for five different randomly extracted learning databases of the same size.

#### IV. RELATED WORK

There exists other feature selection methods also based on neural networks, theoretically described in [7], which we should consider and compare in future works, in the context of intrusion detection. The one we used is the simplest to calculate. We need, thoroughly, to compare the HVS method to other feature selection methods mentioned here, such as SVDF-based method or the one used by [5] based on information gain.

Besides, several recent papers presented various feature selection techniques applied to the KDD features. Reference [8] proposed a hybrid approach combining the information gain ratio (IGR) and the k-means classifier. Reference [9] proposed a feature selection method based on Rough Sets, improved Genetic Algorithms and clustering. Then they used the SVM classifier for performance evaluation on the KDD database. Reference [10] proposed a clustering-based classifier selection method. The method selects the best classifier on similar clusters, compares it with the best classifier on the nearest cluster and chooses the better one to make the system decision. It showed better results than the Clustering and Selection (CS) method.

We should compare our method to these various techniques in a future work. Nevertheless, most of the cited works tested their methods on an extraction from the KDD learning database. They did not test them on the KDD database originally dedicated to testing and containing new attacks as we did in this paper. This demonstrates the potential of the method to detect new attacks and gives more realistic results than the results produced by testing on only a part of the KDD learning database.

#### V. CONCLUSION AND FUTURE WORK

We have shown that the HVS method we presented in this work can be directly and efficiently applied to the problem of intrusion detection, in order to assess the most important features that contribute to attack detection. We could then select a set of most relevant features to accelerate the detection process. An important advantage of the approach, compared to existing methods (like [9]), is that the same technique (feed-forward neural networks) can be used for both feature selection and attack detection, which gives more consistency to the method. Furthermore, the method is almost independent of the used networks' architecture. Further rigorous tests should be conducted to measure accurately the dependence of the HVS method to the learning database, with databases of different sizes. This dependence should not be an obstacle, however, since, in most applications, the learning database is set once for all.

#### ACKNOWLEDGMENT

This work is a partial result of a research project supported by grant number INF 36-8-08 from King Abdul Aziz City for Sciences and Technology (KACST), Riyadh, Kingdom of Saudi Arabia.

#### REFERENCES

- [1]. P.M. Wong, T.D. Gedeon, and I.J. Taggart, "An Improved Technique in Porosity Prediction: A Neural Network Approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33(4), 971-980, 1995.
- [2]. M. Yacoub and Y. Bennani, "HVS: A Heuristic for Variable Selection in Multilayer Artificial Neural Network Classifier", *International Conference on Artificial Neural Networks and Intelligent Engineering, ANNIE '97*, Missouri, USA, 1997, pp. 527-532.
- [3]. The 1998 intrusion detection off-line evaluation plan. MIT Lincoln Lab., Information Systems Technology Group. <http://www.ll.mit.edu/IST/ideval/docs/1998/id98-eval-11.txt>, 25 March 1998.
- [4]. Knowledge discovery in databases DARPA archive. TaskDescription. <http://www.kdd.ics.uci.edu/databases/kddcup99/task.html>
- [5]. H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets," *Proceedings of the Third Annual Conference on Privacy, Security and Trust*, St. Andrews, Canada, October 2005.
- [6]. S. Mukkamala and A. H. Sung, "Identifying Significant Features for Network Forensic Analysis Using Artificial Intelligence Techniques". In the *International Journal on Digital Evidence*, vol. 1(4), 2003.
- [7]. P. Leray and P. Gallinari, "Feature selection with neural networks," *Behaviormetrika*, vol. 26(1), pp. 145-166, 1999.
- [8]. Araújo, N.; de Oliveira, R.; Ferreira, E.-W.; Shinoda, A.A.; Bhargava, B.; , "Identifying important characteristics in the KDD99 intrusion detection dataset by feature selection using a hybrid approach," *Telecommunications (ICT), 2010 IEEE 17th International Conference on*, vol., no., pp.552-558, 4-7 April 2010.
- [9]. Yuteng Guo; Beizhan Wang; Xinxing Zhao; Xiaobiao Xie; Lida Lin; Qingda Zhou; , "Feature selection based on Rough set and modified genetic algorithm for intrusion detection," *Computer Science and Education (ICCSE), 2010 5th International Conference on*, vol., no., pp.1441-1446, 24-27 Aug. 2010.

- [10]. Aizhong Mi and Linpeng Hai, "A clustering-based classifier selection method for network intrusion detection," *Computer Science and Education (ICCSE)*, 2010 5th International Conference on, vol., no., pp.1001-1004, 24-27 Aug. 2010.

# Prospects of high-frequency gravimetry

Alexander L. Dmitriev

**Abstract** - The gravitational field (GF) of the Earth is assumed to be a stochastic process the wide frequency spectrum of which is conditioned by the influence of various geophysical, astrophysical and anthropogenic factors. The frequency range of fluctuations of GF at frequencies over 1 Hz has not been significantly studied yet and still remains a peculiar "Terra Incognita" of gravimetry. Meanwhile, high-frequency changes of a free fall acceleration (FFA) data are informative for understanding of the complex physical processes happening in the core and crust of the Earth. They can be used to solve practical problems such as prediction of earthquakes, exploration of minerals, as well as problems of detection and identification of massive underwater or underground artifacts. Ballistic gravimeters with the test body executed in the form of a mechanical rotor with a horizontal axis of rotation should also be considered as perspective means of HF-gravimetry. Rotary motion corresponds to two oscillatory motions of the rotor particles along the orthogonal axis of coordinates. The accelerated harmonic motion of the rotor particles on a vertical is characterized by an infinite set of time derivatives. In these conditions the interaction of such rotor with a nonstationary gravitational field of Earth can have a specific, not trivial character. Such researches will promote obtaining the new data on dynamic characteristics and specific features of the gravitational field of the Earth.

**Keywords** - ballistic gravimeters, free fall acceleration, gravitational field of the Earth, rotor

## I. INTRODUCTION

The gravitational field of the Earth is assumed to be a stochastic process the wide frequency spectrum of which is conditioned by the influence of various geophysical, astrophysical and anthropogenic factors. High sensitivity of the best modern gravimeters is achieved primarily through proper stabilization of temperature and mechanical characteristic of the equipment used and long integration time of registered signals – from tens of seconds to 24 hours [1]. Obviously, at large times of signal integration, the information about high-frequency variations of a gravitational field is lost. The frequency range of fluctuations  $g_0(t)$  at frequencies over 1 Hz has not been significantly studied yet and still remains a peculiar "Terra Incognita" of gravimetry [2].

Meanwhile, high-frequency changes of a free fall acceleration (FFA) data are informative for understanding of the complex physical processes happening in the core and crust of the Earth.

Alexander L. Dmitriev is with the National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, 49, Kronverksky Prospect, Russia (phone/fax: +7 812 3154071; e-mail: alex@dmitriyev.ru).

They can be used to solve practical problems such as prediction of earthquakes, exploration of minerals, as well as problems of detection and identification of massive underwater or underground artifacts.

High-frequency (HF) gravimetry data is of a great scientific and practical importance and the development of HF-gravimetry as a new research area is inevitable. Such gravimeters should provide an accurate measurement of the "instantaneous" value of FFA in the frequency range from few Hz to thousands (and probably more) Hz.

The most convenient modern tools of HF-gravimetry include superconducting gravimeters (SCG). Owing to a rather big proof mass, the highest frequency of variations in the gravity acceleration value registered by SCG does not exceed a few tens of Hz, although the frequency range of such measurements can be essentially extended after the improvement of these devices. Among HF-gravimetry measurement methods we should also mention the application of ballistic gravimeters with extremely small, less of 1 mm, length of the proof mass fall trajectory [3].

Ballistic gravimeters with the test body executed in the form of a mechanical rotor with a horizontal axis of rotation should also be considered as perspective means of HF-gravimetry.

Rotary motion corresponds to two oscillatory motions of the rotor particles along the orthogonal axis of coordinates. The accelerated harmonic motion of the rotor particles on a vertical is characterized by an infinite set of time derivatives. In these conditions the interaction of such rotor with a nonstationary gravitational field of Earth can have a specific, not trivial character.

## II. WEIGHT OF OSCILLATOR IN A VARIABLE FIELD OF GRAVITATION

Let's consider interaction of a mechanical rotor with an alternating gravitational field which is based on the gravitational analogy of the phenomenon of Faraday and Lenz's Law in electrodynamics [4-6]. According to [5,6] the change of acceleration of the gravity acting on a body, moving with acceleration  $\vec{a}$  under influence of the elastic force, in the elementary (linear) approximation, is represented as

$$\Delta \vec{g}_{p,c} = - \frac{\vec{g}_0}{|\vec{g}_0|} (\vec{g}_0 \cdot \vec{a}) A_{p,c} \quad (1)$$

where symbols  $p,c$  mean passing ( $p$ ) and a contrary ( $c$ ), in relation to a direction of vector  $\vec{g}_0$  of normal acceleration of a gravity, orientation of a

vertical projection of vector  $\vec{a}$  of acceleration of external forces, and factors  $A_p$  and  $A_c$  characterize a degree of change of values  $\Delta\vec{g}_{p,c}$ . If the massive body under action of the external, electromagnetic in nature, elastic force makes harmonious oscillations along a vertical with frequency  $\omega$  and amplitude  $B$ , the average for the period  $\tau = 2\pi/\omega$  of fluctuations value  $\Delta\bar{g}$  of change of FFA of such mechanical oscillator is equal to the sum of average changes of FFA in movement of a body passing and contrary to vector  $\vec{g}_0$ ,

$$\Delta\bar{g} = \Delta\bar{g}_p + \Delta\bar{g}_c \quad (2)$$

and at constant  $g_0 = |\vec{g}_0|$  it is equal

$$\Delta\bar{g} = -\frac{g_0 B \omega^2}{\pi} (A_p - A_c). \quad (3)$$

We shall present elementary time dependence  $g_0(t)$  as

$$g_0(t) = g_0(1 + \beta \sin(\Omega t + \theta)) \quad (4)$$

where  $\Omega$  – frequency of changes of FFA value,  $\beta$  – their relative amplitude,  $\theta$  – the phase. Acceleration  $a(t)$  of the material point making harmonious oscillations along a vertical with amplitude  $B$  is equal to

$$a(t) = B\omega^2 \sin \omega t \quad (5)$$

where  $\omega$  – frequency of oscillations.

The averages for oscillation half-cycle  $\tau/2$  of values of changes of accelerations  $\Delta\bar{g}_p$  and  $\Delta\bar{g}_c$  are equal to

$$\Delta\bar{g}_p = -A_p g_0 B \omega^2 \frac{2}{\tau} \int_0^{\tau/2} \sin \omega t (1 + \beta \sin(\Omega t + \theta)) dt \quad (6)$$

$$\Delta\bar{g}_c = -A_c g_0 B \omega^2 \frac{2}{\tau} \int_{\tau/2}^{\tau} \sin \omega t (1 + \beta \sin(\Omega t + \theta)) dt \quad (7)$$

The relative change of FFA of the oscillator, in view of 2, shall be presented as

$$\frac{\Delta\bar{g}}{g_0} = 4\pi A_p B F^2 f(x) \quad (8)$$

where  $F = \Omega/2\pi$ ,  $x = \omega/\Omega$  and frequency function  $f(x)$  equal to

$$f(x) = -x^2 \left[ \int_0^{\pi} \sin z (1 + \beta \sin(xz + \theta)) dz + \mu \int_{\pi}^{2\pi} \sin z (1 + \beta \sin(xz + \theta)) dz \right] \quad (9)$$

here  $\mu = A_c / A_p$  and  $z = \omega t$ .

Examples of frequency functions  $f(x, \mu, \theta, \beta)$  at various parameters  $\mu, \theta, \beta$ , and both low values of  $x$  are shown in Fig. 1.

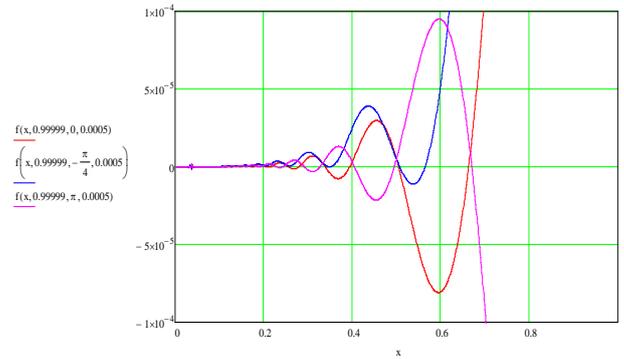


Fig. 1. Frequency functions  $f(x, \mu, \theta, \beta)$  at low values of argument  $x$ ; relative amplitude of fluctuations FFA  $\beta = 0.0005$ .

Obviously, the sign and a general view of functions  $f(x)$  essentially depend on parameters  $\mu, \theta, \beta$ . According to estimations [4,5], in the calculations,  $\mu = 0.99999$  is assumed. The given calculated dependences show that even at small, for example, with relative value of about the 100-th fractions of percent, amplitudes  $\beta$  of fluctuations in value of normal acceleration of the gravity of the Earth, the weight of mechanical oscillator can be changed appreciably.

At frequencies  $\omega$  of oscillations, with an order of the frequency  $\Omega$  of own fluctuations of FFA, in area  $x \leq 1$ , the weight of oscillator is periodically changes with frequency, with sign and values of such changes essentially depending on a difference of phases  $\theta$  of oscillations and FFA (Fig. 1).

### III. EXPERIMENTAL FREQUENCY DEPENDENCE OF FREE FALLING ACCELERATION OF ROTOR

In our experiment the free falling acceleration of the magnetically-, thermally- and sound-isolated container with a vacuumed aviation rotor inside it was measured [7]. Appearance of a rotor is shown in Fig. 2.



Fig. 2. Rotor

The maximal rotation frequency of a rotor is 400 Hz, the run out time of rotor is 22 min. Fall path length of the container is 30 mm, readout time of sample value of gravity acceleration is near 40 ms, the period of sampling is from 0.5 up to 1.0 minutes. The principle of measurements is based on photoregistration of movement of the scale in form of three horizontal strings fixed on the container. At the maximal falling velocity of the container equal to 60 cm/s and its dimensions of 82x82x66 mm, the joint influence of buoyancy and resistance force of air in FFA measurements did not exceed 0.1 cm/s<sup>2</sup>. The error of some measurements of the container FFA was within the limits of 0.3-0.6 cm/s<sup>2</sup> and was basically determined by accuracy of readout times of registration of pulse signals in movement of the scale (near 1 microsecond).

The example of experimental frequency dependence of FFA changes  $\Delta g(f)$  of the container, containing a rotor with a horizontal rotation axis, is shown in the Fig. 2.

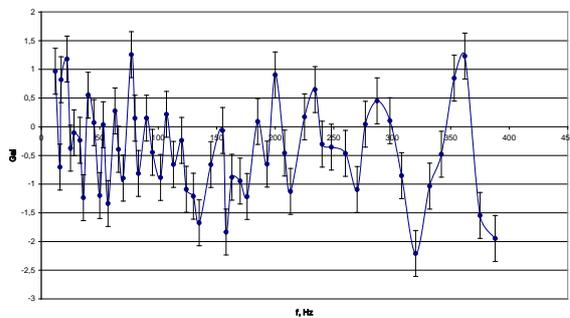


Fig.3. The frequency dependence of free falling acceleration of the container with horizontally positioned rotor; the changes of FFA (*Gal*) relatively to the value of FFA with the stopped rotor have been shown.

The value  $\Delta g(0) = 0$  corresponds to acceleration of free falling of the container with a motionless rotor; FFA measurements of the container with a motionless rotor were carried out till the moment when rotor got going and after its run out time, in so doing the FFA values of the container, averaged by results of 10 measurements with a motionless rotor, coincided to the accuracy of 0.05%.

Comparing Fig. 1 and Fig. 3, it can be seen that the area of steady periodic changes of FFA in Fig. 3 in a band of frequencies 200-400 Hz approximately

corresponds to the area in a vicinity of value  $x \approx 0.5$  in Fig. 1. Having substituted in 8 the experimental value  $\Delta g / g_0 \sim 10^{-3}$ , assume  $A_p \sim 10^{-2} g_0^{-1}$ ,  $f(x) \sim 10^{-5}$ , we obtained an estimation of amplitude  $B \sim 1.4\text{cm}$  of oscillator. The given size almost coincides with radius of the rotor used in experiments. At oscillation frequencies tens times higher than the frequencies  $F$  of own fluctuations of normal acceleration of the gravity (according to the given estimations,  $F \sim 300/0.5 = 600\text{Hz}$ ) and following the suggested model, there is observed a monotonous frequency dependence of change  $\Delta \bar{g}$  of average value of acceleration of free falling oscillator, with sign  $\Delta \bar{g}$  being directly determined by the difference of phases  $\theta$  of fluctuations FFA and oscillator. Within the limits of applicability of formulas 1,5 there are possible both substantial growth and reduction of the average gravity working on mechanical oscillator on the part of the variable gravitational field of the Earth. Let's note that the independent measurements of high-frequency, in the range of hundreds – thousands of Hz, spectra of fluctuations of acceleration of the gravity of the Earth, executed, for example, with use of SCG, will allow to define modes of the matched fluctuations of oscillator at which the changes of its average weight can essentially surpass the ones described by formulas 4-8.

#### IV. CONCLUSION

The calculated and experimental estimations given above have an illustrative character.

Nevertheless, the considered simple phenomenological model finely explains the experimental dependences and agrees with the known data of measurements of weight of accelerated moving test bodies. Experimental researches into free falling mechanical oscillators (rotors, vibrators) will allow to bring the necessary specifications into the offered models, to determine the borders of their applicability, and to prove more strictly the size parameters introduced into these models. Such researches will promote obtaining the new data on dynamic characteristics and specific features of the gravitational field of the Earth. Development of HF-gravimetry techniques and exploration of above-mentioned "Terra Incognita" carries significant scientific and applied value.

#### REFERENCES

- [1] W. Torge, *Gravimetry*, New York: Walter de Gruyter, 2008.
- [2] A. L. Dmitriev, E. M. Nikushchenko, "Prospects and Methods of High-Frequency Gravimetry", *IAG Symposium on Terrestrial Gravimetry (TG-SMM 3013)*, Paper Abstracts, 73-74, Sept.2013.

[3] A. L. Dmitriev, E. I. Kotova et al. “A Ballistic Gravimeter with Dropping Holographic Grating”, *Optics and Spectroscopy*, vol. 117, pp. 799-780, Nov. 2014.

[4] A. L. Dmitriev, “Analogue of Lenz’s Rule in Phenomenological Gravitation” in *AIP Conference Proceedings*, vol. 1103, pp. 345-351, NY 2009.

[5] A. L. Dmitriev, E. M. Nikushchenko and S. A. Bulgakova, “Dynamic Weighing Experiments – the Way to the New Physics of Gravitation”, in *AIP Conference Proceedings*, vol. 1208, pp. 237-246, NY 2010.

[6] A. L. Dmitriev, “Physical Substantiation of an Opportunity of Artificial Change of Body Weight”, *Physics Procedia*, vol. 38, pp.150-163, 2012.

[7] A. L. Dmitriev and E. M. Nikushchenko, “Frequency Dependence of Rotor’s Free Fall Acceleration”, *Engineering Physics*, No 1, 13-17, 2012 (In Russian)

**Prof. Alexander L. Dmitriev** was born in Moscow in 1943. In 1967 he graduated from Dept. of Physics of Leningrad State University. For many years he worked in research and read lectures in the field of physical optics, sensors and lasers. Since 1993 he is a professor at St. Petersburg National Research University of Information Technologies, Mechanics and Optics. Beginning early 90ties, in cooperation with Institute of Metrology was engaged in research of precise weighing. He published more than 100 scientific works include of monographs “Controllable Gravitation”, published in Moscow in 2005, and “Experimental Gravitation”, published in St.-Petersburg in 2014 (in Russian). His main line of research is analogy of optical and gravitational phenomena and experimental gravitation.

# Variable Cosmological Parameter and S-channel Quantum Matter Fields Hadamard renormalization in Spherically Symmetric Curved Space Times

HOSSEIN GHAFARNEJAD\*

**ABSTRACT-** Aim of the paper is to obtain 2d analogue of the backreaction equation which will be useful to study final state of quantum perturbed spherically symmetric curved space times. Thus we take Einstein-massless-scalar  $\psi$  tensor gravity model described on class of spherically symmetric curved space times. We rewrite the action functional in 2d analogue in terms of dimensionless dilaton-matter field ( $\chi = \Phi\psi$ ) where dilaton field  $\Phi$  is conformal factor of 2-sphere. Then we seek renormalized expectation value of quantum dilaton-matter field stress tensor operator by applying Hadamard renormalization prescription. Singularity of the Green function is assumed to be has logarithmic form. Covariantly conservation condition on the renormalized quantum dilaton-matter stress tensor demands to input a variable cosmological parameter  $\lambda(x)$ . Energy conditions (weak, strong and null) is studied on the obtained renormalized stress tensor leading to dynamical equations for  $\lambda(x), \Phi$  and quantum vacuum state  $W_0(x) = \langle 0 | \hat{\chi}^2 | 0 \rangle_{ren}$ . In weak quantum field limits our obtained trace anomaly corresponds to one which obtained from zeta regularization. Setting null-like apparent horizon equation  $\nabla_c \Phi \nabla^c \Phi = 0$ , our procedure predicts that physically correct value of the parameter in the anomaly trace  $\frac{1}{24\pi} \{ R - \alpha \frac{\nabla_c \nabla^c \Phi}{\Phi} + (\alpha - 6) \frac{\nabla_c \Phi \nabla^c \Phi}{\Phi^2} \}$  should be  $\alpha = 6$ .

**Keywords-** Dilaton fields, Dimensional reduction, Hadamard renormalization, Spherically symmetric curved space times, Variable cosmological parameter

## I. INTRODUCTION

In absence of a viable theory of pure quantum gravity, its semiclassical approximation readily yields particle creation in curved background space time (see [1,2] and references therein). In the latter approach the gravitational field is retained as a classical background, while the matter fields are quantized in the usual way. In the latter view the perturbed metric is obtained by the semiclassical Einstein beakreaction equations.

$$G_{\mu\nu} = 8\pi G \{ T_{\mu\nu}^{class} + \langle \hat{T}_{\mu\nu} \rangle_{ren} \} \quad (1)$$

where we used units  $c = \hbar = 1$ ,  $G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R$  with  $\mu, \nu = 0, 1, 2, 3$  is Einstein tensor in four dimensional curved space-time,  $R_{\mu\nu}$  (R) is Ricci tensor

(scalar).  $T_{\mu\nu}^{class}$  is classical matter fields stress tensor.  $\langle \hat{T}_{\mu\nu} \rangle_{ren}$  is renormalized expectation value of quantum matter fields operator. According to wald's axioms [3],  $\langle \hat{T}_{\mu\nu} \rangle_{ren}$  must be covariantly conserved  $\nabla^\mu \langle \hat{T}_{\mu\nu} \rangle_{ren} = 0$ , but in the presence of trace anomaly. For conformally invariant fields the trace anomaly  $\langle \hat{T}_\nu^\nu \rangle_{ren}$  is nonzero, unlike its classical counterpart, and is independent of the quantum state where the expectation value is taken. It is completely expressed in terms of geometrical objects as

$$\langle \hat{T}_\nu^\nu \rangle_{ren} = \frac{1}{2880\pi^2} \{ a C_{\alpha\beta\gamma\delta} C^{\alpha\beta\gamma\delta} + b (R_{\alpha\beta} R^{\alpha\beta} - R^2/3) + c \nabla_\gamma \nabla^\gamma R + d R^2 \} \quad (2)$$

where  $a, b, c, d$  are known as depended on the spin of the quantum fields under consideration [1,2] and  $C_{\alpha\beta\gamma\delta}$  is Weyl tensor. Whether such an approach makes sense is subject to debate. Due to the non-linearity of gravity, it will certainly fail for effects that occur on the scale of the Planck length  $(G\hbar/c^3)^{1/2} = 1.616 \times 10^{-33} cm$ , or involve singularities. Thus it will certainly not be possible to correctly describe, among other things, the very final stage of black holes evaporation in a semiclassical model. On the other hand, one might expect meaningful results as long as one stays in the region exterior of a reasonably sized black hole. It is hoped that the semiclassical approximation in gravity works similarly to the quantum electrodynamics one which is able to describe quantum particles in exterior electromagnetic fields.

Yet in the semiclassical approximation as well as in full quantum gravity, the equations describing the evolution of the system must be solved self-consistently. In four dimensions, this poses a problem: One is only able to calculate the Hawking radiation for a fixed spherically symmetric background metric. Even in the latter case, one obtain instead a relation constraining undetermined function [4] and so study of black hole Hawking radiation in four dimension exhibits with some little success. Hence the Hawking radiation and backreaction effects of created particles on the dynamical background metric is still as an open problem.

In order to get a suitable answer to this problem, one takes two-dimensional analogue of the gravitational models from four dimensions by introducing a dilaton field which contains physical properties of tangential pressure of (classical and quantum) matter fields. The latter idea is a good proposal and zeta function regularization method is used to obtain effective action functionals and corresponding anomaly trace in literature

\*Physics Department, Semnan University, Semnan, IRAN, Zip code: 35131-19111; ghafarnejad@yahoo.com

[5,6,7,8]. In this paper we use other procedure called with Hadamard renormalization prescription. Our procedure inputs a variable cosmological parameter  $\lambda(x)$  reaching to the covariantly conservation condition of the renormalized stress tensor. This variable cosmological parameter is really corrections of an essential effective cosmological constant  $\Lambda_{eff} = \frac{1}{4\pi G}$  defined by Newtonian coupling constant  $G$  which comes from dimensional reduction of space time [9].

Organization of the paper is as follows. In section II we review two dimensional analogue of the Einstein-Hilbert gravity minimally coupled with mass-less scalar field propagating in s-mode. In section III we suggest symmetric two-point Hadamard Green function to be contained logarithmic geometrical singularity. This suggestion is originated from corresponding to Green function of a massless scalar field moving on two dimensional Minkowski flat space time [10]. Hadamard renormalization prescription makes ultraviolet singularities of all physical objects such as, quantum matter action functional, stress tensor expectation value of the field and etc., same as logarithmic geometrical singularity. Renormalized expectation value of the quantum matter stress tensor operator leads to a nonsingular covariantly conserved stress tensor with anomaly trace in the presence of variable cosmological parameter. The suggested variable cosmological parameter is described in terms of derivatives of the dilaton field, Ricci scalar of induced 2d background metric and derivatives of quantum vacuum state  $W_0(x)$ . In section IV we study energy conditions (weak, strong, null) on the obtained renormalized stress tensor which leads to dynamical equations of the fields  $\Phi, W_0(x), \lambda(x)$ . Also our procedure in weak quantum field limits follows results of the one which obtained from zeta function regularization method. In section V we use apparent horizon property of the curved space times on the obtained anomaly trace of quantum matter field stress tensor expectation value. Section VI denotes to concluding remarks.

## II. THE MODEL

We take Einstein-Hilbert gravity interacting with massless scalar matter field  $\psi$  in 4d curved space times

$$I = \frac{1}{16\pi G} \int dx^4 \sqrt{\tilde{g}} \tilde{R} - \frac{1}{2} \int dx^4 \sqrt{\tilde{g}} \tilde{g}^{\mu\nu} \partial_\mu \psi \partial_\nu \psi \quad (3)$$

where  $\tilde{g}$  is absolute value of determinant of the 4d curved space time metric  $\tilde{g}_{\mu\nu}$  ( $\mu, \nu = 0, 1, 2, 3$ ) and  $\tilde{R}$  is its Ricci scalar. Varying the above action with respect to the fields  $\tilde{g}^{\mu\nu}$  and  $\psi$  one can obtain corresponding field equations as

$$\tilde{G}_{\mu\nu} = 8\pi G \tilde{T}_{\mu\nu} \quad (4)$$

and

$$\tilde{\nabla}_\gamma \tilde{\nabla}^\gamma \psi = 0 \quad (5)$$

where

$$\tilde{T}_{\mu\nu} = \partial_\mu \psi \partial_\nu \psi - \frac{1}{2} \tilde{g}_{\mu\nu} \{ \partial_\gamma \psi \partial^\gamma \psi \} \quad (6)$$

in which Bianchi identity  $\tilde{\nabla}^\mu \tilde{G}_{\mu\nu} = 0$  leads to covariant conservation condition of the matter field

$$\tilde{\nabla}^\mu \tilde{T}_{\mu\nu} = 0, \quad \tilde{T}_\mu^\mu = -\tilde{g}^{\mu\nu} \partial_\mu \psi \partial_\nu \psi. \quad (7)$$

We choose class of 4d spherically symmetric curved space times metrics as

$$ds^2 = \tilde{g}_{\mu\nu} dx^\mu dx^\nu \\ = g_{ab}(x^a) dx^a dx^b + \Phi^2(x^a) (d\theta^2 + \sin^2 \theta d\phi^2) \quad (8)$$

where signature of the metric (8) is assumed to be  $(-, +, +, +)$ . Then we assume that the metric fields  $g_{ab}$ ,  $\Phi$  and matter field  $\psi$  are independent of angular coordinates  $(\theta, \phi)$  propagating in spherically modes (S-channel) and integrate (3) with respect to angular coordinates  $\theta$  and  $\phi$  leading to [9]

$$I = \frac{1}{4G} \int dx^2 \sqrt{g} \{ 1 + g^{ab} \partial_a \Phi \partial_b \Phi + \frac{1}{2} \Phi^2 R \} \\ - 2\pi \int dx^2 \sqrt{g} \Phi^2 \partial_a \psi \partial^a \psi. \quad (9)$$

$\Phi$  is called geometrical dilaton field with *length* dimensions and it is in agreement with the status of boson particles in point of view of field theory.  $g_{ab}(x^0, x^1)$  is 2d induced metric on the hypersurface  $\theta = \phi = \text{constant}$ .  $g$  is absolute value of determinant of 2d metric  $g_{ab}$  and  $R$  is corresponding 2d Ricci scalar. The matter field  $\psi$  has inverse of length dimensions. Varying (9), with respect to  $g_{ab}$ ,  $\Phi$ , and  $\psi$ , the corresponding field equations are obtained respectively as [9]

$$\Phi^2 \tilde{G}_{ab} = -2\Phi \nabla_a \nabla_b \Phi + g_{ab} \{ 2\Phi \nabla_c \nabla^c \Phi + \partial_c \Phi \partial^c \Phi - 1 \} \\ = 8\pi G \Phi^2 \tilde{T}_{ab}[\psi], \quad (10)$$

$$\tilde{G}_{\theta\theta} = \Phi \nabla_c \nabla^c \Phi - \frac{1}{2} R \Phi^2 = 8\pi G \tilde{T}_{\theta\theta} = -4\pi G \Phi^2 \partial_c \psi \partial^c \psi, \quad (11)$$

and

$$\nabla_c \nabla^c \psi = -2J^a \nabla_a \psi, \quad J_a = \nabla_a \ln \Phi \quad (12)$$

where we defined

$$\nabla_a \nabla^a = \frac{1}{\sqrt{g}} \partial_a (\sqrt{g} g^{ab} \partial_b), \quad \partial_a \equiv \frac{\partial}{\partial x^a} \quad (13)$$

and non-angular components of the stress tensor (6) as

$$\tilde{T}_{ab}[\psi] = \partial_a \psi \partial_b \psi - \frac{1}{2} g_{ab} \partial_c \psi \partial^c \psi. \quad (14)$$

The matter stress tensor (14) is trace free but same as (7) dose not satisfy the covariant conservation condition in 2d space times. Applying (13) and (14) one can obtain

$$\nabla^a T_{ab} = -2J_a \partial^a \psi \partial_b \psi, \quad T_a^a = 0 \quad (15)$$

where we are dropped over tilde  $\sim$ . Violation of covariant conservation is caused because of non-vanishing dilaton current  $J_a$  and it is coupled with matter current  $\partial_a \psi$  as a source in RHS of the matter wave equation (12). The quantity  $J_a \partial^a \psi$  treats as scalar charge for the field  $\psi$  from view of string theory. Originally this charge comes from dynamical effects of reference frames. For instance in higher dimensional string theory of gravity the Brans-Dicke scalar tensor theory is charge-less and so a covariantly conserved model in Jordan frame but it is not in other frames (see Ref. [13] chapter 2). Hence the string theory accepts that the Bianchi identity no longer implies the covariant conservation of the stress tensors separately in 4+d dimensional curved space times. In other words stress tensors of matter and geometrical dilaton fields do not need follow covariant conservation conditions separately. Physically non-conservation condition of the stress tensor implies that the motion of a free test particle is no longer geodesic when the particle has an intrinsic scalar charge and the gravitational background contains a non-trivial dilaton component.

However we follow here other point of view: dimensional reduction of the space times causes to break covariant conservation condition. On the other hand we know that renormalization of the quantum matter fields breaks also the covariant conservation condition of the stress tensor (see [1,2] and references therein). Some applicable methods are presented to satisfy the covariant conservation condition but by inducing anomaly trace. What is correspondence between them to obtain both quantum matter stress tensor and its geometrical classical dilaton counter part satisfying covariantly conservation condition separately in 2d gravity model (9)? In the following section we try to obtain a suitable answer to this question. We apply Hadamard renormalization prescription to evaluate regular expectation value of quantum matter stress tensor operator  $\langle \hat{T}_{ab}[\hat{\psi}] \rangle_{ren}$  by presenting a variable cosmological parameter  $\lambda(x)$ .

### III. HADAMARD RENORMALIZATION

If  $\psi$  treats as massless quantum bosons. Then it will be linear operator  $\hat{\psi}$  operating on arbitrary state of Hilbert space. Corresponding stress energy tensor operator  $\hat{T}_{ab}[\hat{\psi}]$  become bi-linear with respect to  $\hat{\psi}$  and regular stress tensor counterpart  $\langle \hat{T}_{ab}[\hat{\psi}] \rangle_{ren}$  (subscript 'ren' denotes to 'renormalized') is obtained by eliminating its ultraviolet divergence terms, in one loop level. With given  $\langle \hat{T}_{ab}[\hat{\psi}] \rangle_{ren}$  one can write two dimensional analogue of the metric back reaction equation (1) by re-

garding (10) and (11) such as follows.

$$\begin{aligned} G_{ab} &= -\frac{2\nabla_a \nabla_b \Phi}{\Phi} + g_{ab} \left\{ \frac{2\nabla_c \nabla^c \Phi}{\Phi} + \frac{\partial_c \Phi \partial^c \Phi}{\Phi^2} - \frac{1}{\Phi^2} \right\} \\ &= 8\pi G \frac{\langle \Phi^2 \hat{T}_{ab}[\hat{\psi}] \rangle_{ren}}{\Phi^2} \end{aligned} \quad (16)$$

and

$$G_{\theta\theta} = \Phi \nabla_c \nabla^c \Phi - \frac{1}{2} R \Phi^2 = -4\pi G \langle \Phi^2 \partial_c \hat{\psi} \partial^c \hat{\psi} \rangle_{ren} \quad (17)$$

where  $g_{ab}$  and  $\Phi$  is still treats as classical geometrical fields whereas the matter field  $\psi$  is assumed to be treat as quantum field. Furthermore we would not move  $\Phi$  outside the expectation quantities  $\langle \Phi^2 \hat{T}_{ab}[\hat{\psi}] \rangle_{ren}$  and  $\langle \Phi^2 \partial_c \hat{\psi} \partial^c \hat{\psi} \rangle_{ren}$ , because variable dilaton field causes to violation of covariantly conservation of matter stress tensor  $T_{ab}[\psi]$  in its classical regime (see Eq. (15)). Applying (16) and (17) the Bianchi identity  $\nabla^\mu G_{\mu\nu} = 0$  in 4d leads to the following constraint condition.

$$\nabla^a \langle \Phi^2 \hat{T}_{ab}[\hat{\psi}] \rangle_{ren} = \nabla_b \left( \frac{1}{\Phi^2} \right) \langle \Phi^2 \partial_c \hat{\psi} \partial^c \hat{\psi} \rangle_{ren}. \quad (18)$$

In 4d space times  $\Phi$  and  $1/\psi$  has length dimension and conformal invariance property of the matter action in (3) is broken in 2d analogue (9). Hence it will be useful we define a dimensionless dilaton-matter field as

$$\chi = \Phi \psi \quad (19)$$

before than that we proceed to apply renormalization prescription and evaluate expectation value of its stress tensor operator  $\langle \hat{T}_{ab}[\hat{\chi}] \rangle_{ren}$ . Applying (19), one can rewrite matter part of the action (9) as

$$\begin{aligned} I_{matter}[\chi, g_{ab}, \Phi] &= 2\pi \int \sqrt{g} dx^2 g^{ab} \{ \nabla_a \chi \nabla_b \chi + \chi^2 J_a J_b \\ &\quad - \chi J_b \nabla_a \chi - \chi J_a \nabla_b \chi \}. \end{aligned} \quad (20)$$

Dynamical equation of the field  $\chi$  is obtained by varying the above action with respect to  $\chi$  as

$$\left\{ \nabla_c \nabla^c - \frac{\nabla_c \nabla^c \Phi}{\Phi} \right\} \chi = 0 \quad (21)$$

Trace free  $T_a^a[\chi]$  stress tensor of the field  $\chi$  is obtained by varying (20) with respect to  $g^{ab}$  such as follows.

$$\begin{aligned} T_{ab}[\chi] &= \nabla_a \chi \nabla_b \chi + \chi^2 J_a J_b - \chi (J_a \nabla_b \chi + J_b \nabla_a \chi) \\ &\quad - \frac{g_{ab}}{2} \{ \nabla_c \chi \nabla^c \chi + \chi^2 J_c J^c - 2\chi J_c \nabla^c \chi \} \end{aligned} \quad (22)$$

which is equivalent with  $\Phi^2 T_{ab}[\psi]$  and so we can deduce

$$\langle \Phi^2 \hat{T}_{ab}[\hat{\psi}] \rangle_{ren} \equiv \langle \hat{T}_{ab}[\hat{\chi}] \rangle \quad (23)$$

and

$$\langle \Phi^2 \partial_c \hat{\psi} \partial^c \hat{\psi} \rangle \equiv$$

$$\langle \partial^c \hat{\chi} \partial_c \hat{\chi} \rangle - 2J^c \langle \hat{\chi} \partial_c \hat{\chi} \rangle + J_c J^c \langle \hat{\chi}^2 \rangle. \quad (24)$$

In the following we seek renormalized expectation values of the quantities (22) and (24) by applying the Hadamard renormalization prescription.

This approach is begun with definition of the expectation value of stress tensor (22) such as follows.

$$\langle \hat{T}_{ab}[\hat{\chi}] \rangle = \lim_{x' \rightarrow x} D_{ab}(x, x') G^+(x, x') \quad (25)$$

where a state of  $\hat{\chi}$  is characterized by a hierarchy of Wightman function which for a symmetric two-point function we have

$$G^+(x, x') = \frac{1}{2} \langle \hat{\chi}(x) \hat{\chi}(x') + \hat{\chi}(x') \hat{\chi}(x) \rangle \quad (26)$$

and

$$\begin{aligned} D_{ab}(x, x') &= g_b^b \nabla_a \nabla_{b'} + g_a^a \nabla_{a'} \nabla_b + J_a J_b \\ &\quad - J_a \{ \nabla_b + g_b^b \nabla_{b'} \} - J_b \{ \nabla_a + g_a^a \nabla_{a'} \} \\ &\quad - g^{ab} \{ g_c^c \nabla_c \nabla_{c'} - J_c (\nabla^c + \nabla^{c'}) + \frac{J_c J^c}{2} \} \end{aligned} \quad (27)$$

with the bivector of parallel transport  $g_a^a$ , is the bilocal differential operator. This expression makes explicit that the singular character of the operator  $\hat{T}_{ab}$  emerges as a consequence of the short-distance singularity of the symmetric two-point function  $G^+(x, x')$ . Equivalence principle suggest that the leading singularity of  $G^+(x, x')$  should have a close correspondence to singularity structure of the two-point function of massless fields in Minkowski space [10]. In general the entire singularity of  $G^+(x, x')$  may have a more complicated structure. Usually one assumes that  $G^+(x, x')$  has a singular structure represented by the Hadamard expansions. This means that in a normal neighborhood of a point  $x$  in 2d curved space time, we can suggest logarithmic dependence (Hadamard Green functions in 4d curved space times have singularities same as  $\sigma^{-1}$  and  $\ln \sigma$  [1,2,14,15,16].) of the Green function  $G^+(x, x')$  for a massless quantum scalar field  $\chi$  as

$$G^+(x, x') = V(x, x') \ln \sigma(x, x') + W(x, x') \quad (28)$$

where  $2\sigma(x, x') = \sigma^a \sigma_a$  with  $\sigma_a \equiv \nabla_a \sigma$ , is one-half square of the geodesic distance between  $x$  and  $x'$ . Non-singular two point functions  $V(x, x')$ , and  $W(x, x')$  have the following power series expansions

$$V(x, x') = \sum_{n=0}^{\infty} V_n(x, x') \sigma^n \quad (29)$$

and

$$W(x, x') = \sum_{n=0}^{\infty} W_n(x, x') \sigma^n \quad (30)$$

where  $V(x, x')$  ( $W(x, x')$ ) is state-independent (dependent) 2 point functions. The Green function (28) satisfies the field equation (21) with respect to both points  $x$  and  $x'$  as

$$\{ \nabla_c \nabla^c - \frac{\nabla_c \nabla^c \Phi}{\Phi} \} G^+(x, x') = g^{-1/2} \delta^2(x - x') \quad (31)$$

where  $\delta^2(x - x')$  is well known Dirac delta function in 2 dimensions. Applying (28), (29), (30) and (31), with  $x' \neq x$ , the coefficients  $V_n(x, x')$  and  $W_n(x, x')$  satisfies the following recursion relations.

$$2(n+1)^2 V_{n+1} + 2(n+1) \nabla_a V_{n+1} \sigma^a + (\nabla_c \nabla^c - \frac{\nabla_c \nabla^c \Phi}{\Phi}) V_n = 0, \quad (32)$$

$$\begin{aligned} (\nabla_c \nabla^c - \frac{\nabla_c \nabla^c \Phi}{\Phi}) W_n + 2(n+1) \nabla_a W_{n+1} \sigma^a + 2(n+1)^2 W_{n+1} \\ + 4(n+1) V_{n+1} + 2 \nabla_a V_{n+1} \sigma^a = 0. \end{aligned} \quad (33)$$

Covariant Taylor series expansion for symmetric two point functions is written as [14,15] (see also [16])

$$\begin{aligned} \Gamma(x, x') &= \Gamma(x) - \frac{1}{2} \nabla_a \Gamma(x) \sigma^a + \frac{1}{2} \Gamma_{ab}(x) \sigma^a \sigma^b \\ &\quad + \frac{1}{4} \{ \frac{1}{6} \nabla_c \nabla_b \nabla_a \Gamma(x) - \nabla_c \Gamma_{ab}(x) \} \sigma^a \sigma^b \sigma^c + O(\sigma^2) \end{aligned} \quad (34)$$

which for  $W(x, x')$  we obtain from coincidence limits

$$\begin{aligned} W(x) &= \lim_{x' \rightarrow x} W(x, x') = \lim_{x' \rightarrow x} W_0(x, x') \\ &= W_0(x) = \langle \hat{\chi}^2 \rangle_{ren} = \langle \Phi^2 \hat{\psi}^2 \rangle_{ren}. \end{aligned} \quad (35)$$

The above renormalized expectation value is called vacuum state of the quantum dilaton-matter field  $\chi$ . Also one can obtain from coincidence limits of the equations (32), (33), (34) and (35)

$$V_1(x) = \frac{1}{2} \left\{ \frac{\nabla_c \nabla^c \Phi}{\Phi} V_0(x) - V_{0c}^c(x) \right\}, \quad (36)$$

$$W_1(x) = V_{0c}^c(x) - \frac{W_{0c}^c(x)}{2} + \left( \frac{W_0(x)}{2} - V_0(x) \right) \frac{\nabla_c \nabla^c \Phi}{\Phi} \quad (37)$$

and

$$W_{ab}(x) = W_{0ab}(x) + W_1(x) g_{ab}. \quad (38)$$

Applying (28) and (34) for  $\Gamma(x, x') = V_0(x, x')$  the equation (31) with  $x' \neq x$  leads to

$$\left\{ \nabla_c \nabla^c - \frac{\nabla_c \nabla^c \Phi}{\Phi} \right\} W(x, x') = \frac{V_0(x)}{3} \left\{ R_{ab} \frac{\sigma^a \sigma^b}{\sigma} - \frac{1}{4} \nabla_a R_{bc} \frac{\sigma^a \sigma^b \sigma^c}{\sigma} \right\} + O(\sigma) \quad (39)$$

Inserting (34) for  $\Gamma(x, x') = W(x, x')$ , the above equation reduces to the following conditions.

$$W_c^c(x) = \frac{\nabla_c \nabla^c \Phi}{\Phi} W_0(x) + \frac{V_0(x)}{3} R \quad (40)$$

and

$$\begin{aligned} \nabla^b \left[ 3\widetilde{W}_{0ab}(x) + \frac{g_{ab}}{4} (V_0(x)R - 3\nabla_c \nabla^c W_0(x) - \lambda(x)) \right] \\ = R_{ae} \nabla^e W_0(x) \end{aligned} \quad (41)$$

where

$$\widetilde{W}_{0ab}(x) = W_{0ab}(x) - \frac{1}{2} g_{ab} W_{0c}^c(x) \quad (42)$$

and we used identities

$$\nabla_c \nabla^c \nabla^b W_0(x) = \nabla^b \nabla_c \nabla^c W_0(x) + R^{ab} \nabla_a W_0(x), \quad (43)$$

$$\nabla^b \nabla_a \nabla_b W_0(x) = \nabla_a \nabla_c \nabla^c W_0(x) + R_{ab} \nabla^b W_0(x). \quad (44)$$

We defined ‘effective variable cosmological parameter’  $\lambda(x)$  satisfying the constraint condition

$$R \nabla_a V_0(x) = \nabla_a \lambda(x) \quad (45)$$

and also applied

$$V_{0a}^a(x) = V_0(x) \left( \frac{R}{6} + \frac{\nabla_c \nabla^c \Phi}{\Phi} \right), \quad V_1(x) = -\frac{V_0(x)R}{12} \quad (46)$$

$$W_{ab}(x) = \widetilde{W}_{0ab}(x) + g_{ab} \left( \frac{V_0(x)R}{6} + \frac{W_0(x)}{2} \frac{\nabla_c \nabla^c \Phi}{\Phi} \right) \quad (47)$$

which are obtained from (36), (37), (38), (40). Now we subtract from  $G^+(x, x')$  defined by (28), a local symmetric two point function  $G_L^+(x, x')$  with the same short-distance singularity of the Hadamard expansion. Then we make a renormalized expectation value of stress tensor (25) as

$$\langle \hat{T}^{ab}[\hat{\chi}] \rangle_{ren} = \lim_{x' \rightarrow x} D^{ab}(x, x') \{ G^+(x, x') - G_L^+(x, x') \} \quad (48)$$

which by applying (28) can be rewritten as

$$\langle \hat{T}^{ab}[\hat{\chi}] \rangle_{ren} = \lim_{x' \rightarrow x} D^{ab}(x, x') \{ W(x, x') \}. \quad (49)$$

Explicit form of the nonsingular stress tensor (49) is obtained by inserting (34) [with  $\Gamma(x, x') = W(x, x')$ ], (47) and taking its coincidence limit as

$$\langle \hat{T}_{ab}[\hat{\chi}] \rangle_{ren} = \nabla_a \nabla_b W_0(x) - 2\widetilde{W}_{0ab}(x) -$$

$$\frac{3}{2} (J_a \nabla_b + J_b \nabla_a) W_0(x) + J_a J_b W_0(x) +$$

$$g_{ab} \left\{ J_c \nabla^c W_0(x) - \frac{\nabla_c \nabla^c W_0(x)}{2} - \frac{J^c J_c}{2} W_0(x) \right\} \quad (50)$$

where  $\langle \hat{T}_a^a[\hat{\chi}] \rangle_{ren} = -J_c \nabla^c W_0(x)$ . With same calculation one can obtain for (24):

$$\langle g^{ab} \nabla_a \hat{\chi} \nabla_b \hat{\chi} \rangle_{ren} = \lim_{x' \rightarrow x} D(x, x') \{ W(x, x') \} =$$

$$\frac{\nabla_c \nabla^c W_0(x)}{2} - J_c \nabla^c W_0(x) + J_c J^c W_0(x) -$$

$$\frac{V_0(x)}{3} R + W_0(x) \frac{\nabla_c \nabla^c \Phi}{\Phi} \quad (51)$$

where we defined

$$D(x, x') = g_a^a \nabla^a \nabla_a'. \quad (52)$$

Applying (50) and identity (43) one can obtain

$$\nabla^b \left\{ \langle \hat{T}_{ab}[\hat{\chi}] \rangle_{ren} + 2\widetilde{W}_{0ab}(x) + \frac{3}{2} (J_a \nabla_b + J_b \nabla_a) W_0(x) \right.$$

$$\left. - J_a J_b W_0(x) + g_{ab} \left( \frac{1}{2} J_c J^c W_0(x) - J_c \nabla^c W_0(x) \right. \right.$$

$$\left. \left. - \frac{1}{2} \nabla_c \nabla^c W_0(x) \right) \right\} = R_{ae} \nabla^e W_0(x). \quad (53)$$

Subtracting (41) from (53) we obtain

$$\nabla^a \Sigma_{ab} = 0 \quad (54)$$

where  $\Sigma_{ab}$  is general state independent divergence-less stress tensor relating to  $\langle \hat{T}_{ab}[\hat{\chi}] \rangle_{ren}$  as

$$\langle \hat{T}_{ab}[\hat{\chi}] \rangle_{ren} = -\Sigma_{ab} + \widetilde{W}_{0ab}(x) + J_a J_b W_0(x) -$$

$$\frac{3}{2} (J_a \nabla_b + J_b \nabla_a) W_0(x) + g_{ab} \left\{ \frac{V_0(x)R}{4} -$$

$$\frac{\lambda(x)}{4} - \frac{\nabla_c \nabla^c W_0(x)}{4} - \frac{J_c J^c W_0(x)}{2} + J_c \nabla^c W_0(x) \right\} \quad (55)$$

with

$$\langle \hat{T}_a^a[\hat{\chi}] \rangle_{ren} = -\Sigma_a^a + \frac{V_0(x)R}{2} - \frac{\lambda(x)}{2}$$

$$-J_c \nabla^c W_0(x) - \frac{\nabla_c \nabla^c W_0(x)}{2}. \quad (56)$$

The stress tensor  $\Sigma_{ab}$  is really geometric counterpart of the back reaction equation (16) in 2d analogue and other terms in (55) denotes to matter dependent counter part. This is subject which we seek to answer the question presented in last paragraph of the section 2 of the paper. Inserting (51) into RHS of the equation (17) one can obtain

$$\begin{aligned} \nabla_c \nabla^c W_0(x) - 2J_c \nabla^c W_0(x) + 2 \left( J_c J^c + \frac{\nabla_c \nabla^c \Phi}{\Phi} \right) W_0(x) \\ = \left( \frac{2V_0(x)}{3} + \frac{\Phi^2}{4\pi G} \right) R - \frac{\Phi \nabla_c \nabla^c \Phi}{2\pi G}. \end{aligned} \quad (57)$$

Applying (56) and trace of the equation (16) we obtain

$$\begin{aligned} \Sigma_c^c = \frac{1}{4\pi G} - \frac{\lambda(x)}{2} + \frac{V_0(x)R}{2} - \frac{\Phi \nabla_c \nabla^c \Phi}{4\pi G} - \frac{\Phi^2 J_c J^c}{4\pi G} \\ - J_c \nabla^c W_0(x) - \frac{\nabla_c \nabla^c W_0(x)}{2}. \end{aligned} \quad (58)$$

Applying the above relation and (55) the backreaction equation (16) reduces to

$$\begin{aligned} \tilde{\Sigma}_{ab} - \frac{1}{4\pi G} \left\{ \Phi \nabla_a \nabla_b \Phi - \frac{1}{2} g_{ab} \Phi \nabla_c \nabla^c \Phi \right\} = \\ \tilde{W}_{0ab}(x) + W_0(x) \left[ J_a J_b - \frac{1}{2} g_{ab} J_c J^c \right] \\ - \frac{3}{2} \left[ (J_a \nabla_b + J_b \nabla_a) W_0(x) - g_{ab} J_c \nabla^c W_0(x) \right] \end{aligned} \quad (59)$$

where defined

$$\tilde{\Sigma}_{ab}(x) = \Sigma_{ab}(x) - \frac{1}{2} g_{ab} \Sigma_c^c(x). \quad (60)$$

Applying (51), (55), (58) and (59) the Bianchi identity (18) leads to the following constraint condition.

$$\begin{aligned} V_0(x) = \frac{3\Phi^2}{4\pi G} - \frac{3\Phi \nabla_c \nabla^c \Phi}{2\pi GR} + \\ \frac{3\Phi^2 J^b \nabla^a [\Phi^2 J_c J^c + 5\Phi \nabla_c \nabla^c \Phi - 2\Phi \nabla_a \nabla_b \Phi]}{8\pi GR (J_c J^c)}. \end{aligned} \quad (61)$$

Applying the above result one can obtain explicit form of the cosmological parameter  $\lambda(x)$  from (45) as

$$\lambda(x) = \int R(x) \nabla_a V_0(x) dx^a + Constant. \quad (62)$$

This equation denotes to fluctuations of the variable cosmological parameter  $\lambda(x)$  satisfying to the wave equation

$$\nabla_c \nabla^c \lambda(x) - \nabla^c \ln R \nabla_c \lambda = R \nabla_c \nabla^c V_0(x). \quad (63)$$

This wave equation is derived from constraint condition (45) and its RHS treats as geometrical source.

However for a fixed 2d background metric  $g_{ab} dx^a dx^b$ , we obtained 6 equations defined by (54), (57), (58), (59), (61) and (62) which are not enough to determine seven quantities  $W_0(x)$ ,  $\lambda(x)$ ,  $V_0(x)$ ,  $\Phi(x)$ ,  $\Sigma_{ab}$ ,  $\Sigma_c^c$  and  $W_{0ab}(x)$ . Explicit form of all these quantities are depended to form of the dilaton field  $\Phi$ . What is its dynamical equation? In particular spherically symmetric static space times with  $\Phi(r) = r$  one can continue to solve the above equations and obtain the foregoing dynamical fields but this is a bad restriction on our procedure. For general dynamical 4d spherically symmetric curved space times we should have other management. Usually energy conditions play important role on the physical sources. We study energy conditions on 4d counter part of quantum matter stress tensor given by (51), (55) and (56) in the following section.

#### IV. ENERGY CONDITIONS

In general we consider time-like curves whose tangent 4-vector  $V^\mu = (V^a, 0, 0)$ , with  $V^\mu V_\mu = \beta > 0$ ,  $a = 0, 1$  and background metric signature  $(-, +, +, +)$  which represents the radial velocity vector of a family observer. In the latter case weak (WEC) and strong (SEC) energy conditions leads to

$$WEC : \quad \langle \Phi^2 \hat{T}_{ab}[\hat{\psi}] \rangle_{ren} V^a V^b = \eta \geq 0 \quad (64)$$

and

$$SEC : \quad \langle \Phi^2 \hat{T}_{ab}[\hat{\psi}] \rangle_{ren} V^a V^b -$$

$$\frac{1}{2} \{ \langle \Phi^2 \hat{T}_a^a[\hat{\psi}] \rangle_{ren} - \langle \Phi^2 \partial_c \hat{\psi} \partial^c \hat{\psi} \rangle_{ren} \} V^a V_a = \delta \geq 0. \quad (65)$$

There is also a null energy condition (NEC) for radial null vector field  $N^\mu = (N^a, 0, 0)$  with  $N^\mu N_\mu = 0$  and  $a = 0, 1$  as

$$NEC : \quad \langle \Phi^2 \hat{T}_{ab}[\hat{\psi}] \rangle_{ren} N^a N^b = \sigma \geq 0. \quad (66)$$

Obviously, the above energy conditions emerge directly from the geodesic structure of the spherically symmetric space time (8).

Defining

$$V^a J_a = \alpha, \quad V^a V_a = \beta > 0, \quad N^a J_a = \gamma \quad (67)$$

and applying (51), (55), and (56) the energy conditions (64), (65) and (66) leads to the following relations respectively.

$$WEC : \quad (W_{0ab} - \Sigma_{ab}) V^a V^b + W_0(x) \left( \alpha^2 - \frac{\beta}{2} J^c J_c \right) +$$

$$(\beta J_c - 3\alpha V_c)\nabla^c W_0(x) +$$

$$\frac{\beta}{4} \left( V_0(x)R - \lambda - \nabla_c \nabla^c W_0(x) - 2W_{0c}^c(x) \right) = \eta \quad (68)$$

$$SEC: \quad \Sigma_c^c = \frac{2(\delta - \eta)}{\beta} - \frac{\lambda(x)}{2} + \frac{5V_0(x)}{6} R - \nabla_c \nabla^c W_0(x) -$$

$$\left( J_c J^c + \frac{\nabla_c \nabla^c \Phi}{\Phi} \right) W_0(x) \quad (69)$$

and

$$NEC: \quad (W_{0ab} - \Sigma_{ab})N^a N^b + \gamma^2 W_0(x) -$$

$$3\gamma N^c \nabla_c W_0(x) = \sigma. \quad (70)$$

Applying (57) and (58), the SEC given by (69) leads to the following wave equation.

$$\nabla_c \nabla^c \Phi^2 - \left( J_c J^c + \frac{R}{2} \right) \Phi^2 = 1 + \frac{8\pi G(\eta - \delta)}{\beta} \quad (71)$$

where we used identity  $2\Phi \nabla_c \nabla^c \Phi + 2\Phi^2 J_c J^c = \nabla_c \nabla^c \Phi^2$ . This equation describes evolutions of surface area of apparent horizon  $S = 4\pi\Phi^2$  of the 4d spherically symmetric space time (7) propagating in 2d induced space time  $g_{ab}dx^a dx^b$ . With (71), our strategy about formulation of 2d analogue of the backreaction equation (1) and the renormalized expectation value of the quantum matter-dilaton field stress tensor operator is finished. It will be useful now we imply apparent horizon property of the 4d spherically symmetric curved space time (8) on our derived equations.

## V. APPARENT HORIZON

Assuming  $S = 4\pi\Phi^2$  to be surface area of apparent horizon of the spherically symmetric curved space time (8), one can obtain its position by the null condition

$$g^{ab} \nabla_a S \nabla_b S = 0 \quad (72)$$

which by defining  $J_a = \nabla_a \ln \Phi$  leads to the condition

$$J_a J^a = 0. \quad (73)$$

In this case we can use  $J_a = N_a$  as a suitable null vector field in the NEC (66) for which  $\gamma = 0$  (see (67)). In this case the NEC given by (70) leads to

$$(W_{0ab}(x) - \Sigma_{ab})J^a J^b = \sigma \geq 0. \quad (74)$$

Setting  $\sigma = 0$  we can choose

$$W_{0ab}(x) = \Sigma_{ab}(x) + \xi g_{ab} \quad (75)$$

where  $\xi$  is arbitrary constant parameter. Using (73) and (75) the WEC (68) and SEC (69) leads to respectively

$$WEC: \quad W_{0c}^c(x) = 2\xi + \frac{V_0(x)R}{2} - \frac{\lambda(x)}{2} - \frac{\nabla_c \nabla^c W_0(x)}{2} + \frac{2}{\beta} \{ \alpha^2 W_0(x) + (\beta J_c - 3\alpha V_c) \nabla^c W_0(x) - \eta \} \quad (76)$$

and

$$SEC: \quad \Sigma_c^c = \frac{2(\delta - \eta)}{\beta} - \frac{\lambda(x)}{2} + \frac{5V_0(x)}{6} R - \nabla_c \nabla^c W_0(x) - \frac{\nabla_c \nabla^c \Phi}{\Phi} W_0(x). \quad (77)$$

One of trivial solutions of the equation (45) is slow varying regime of the cosmological parameter  $\lambda(x)$  for which we can exclude its derivatives as

$$\lambda(x) = \frac{4(\delta - \eta)}{\beta} \cong \text{constant}, \quad V_0(x) = \frac{1}{20\pi}. \quad (78)$$

Under the latter assumptions the anomaly trace (5.6) become

$$\Sigma_c^c \cong \frac{R}{24\pi} - \omega \frac{\nabla_c \nabla^c \Phi}{\Phi} \quad (79)$$

in weak quantum field (WQF) limits as

$$W_0(x) \approx \text{constant} = \omega > 0 \quad (80)$$

by excluding its derivatives. The anomaly trace (79) follows well known one which is derived from zeta function regularization method in 2d dilaton quantum field theory [5,6,7,8,17,18,19,20,21,22,23,24,25,26,27,28,29] as

$$\Sigma_c^c(x) = \frac{1}{24\pi} \left\{ R - \alpha \frac{\nabla_c \nabla^c \Phi}{\Phi} + (\alpha - 6) \frac{\nabla_c \Phi \nabla^c \Phi}{\Phi^2} \right\} \quad (81)$$

The arbitrary parameter  $\alpha$  is the coefficient in question [29].  $\alpha = -2$  proposed by R. Bousso and S. W. Hawking [17] which turned out to be a mistake.  $\alpha = 4$  obtained by Kummer et al. [7,18,19] for the same setup of the two-dimensional model as was used by Bousso and Hawking.  $\alpha = 6$  obtained by Elizalde et al. [20] and V. Mukhanov, A. Wipf and A. Zelnikov [5]. This result turned out to be correct physically satisfying our statement about apparent horizon induction on the anomaly. In other word (81) reduces to (79) by setting

$$\alpha = 6, \quad \omega = \frac{1}{4\pi}. \quad (82)$$

In strong quantum field limits where we can not exclude fluctuations of the field  $W_0(x)$  and so its derivatives should be considered in procedure one should be follow exact equations given in the previous section. In general, our procedure is useful to study final state of quantum

perturbed 4d spherically symmetric curved space times. Asymptotically flat classical static metric solution of the model (3) was obtained previously by Jains-Newman-Winicour (JNW) [11,12]. As a future work one can use the presented formalism to study physical effect of the obtained anomaly on the quantum perturbed JNW metric solution.

## VI. CONCLUDING REMARK

In this article we used 2d analogue of the Einstein-massless scalar gravity to study 4d spherically symmetric quantum field theory. Hadamard renormalization prescription is used to obtain renormalized matter-dilaton stress tensor in the presence of variable cosmological parameter which has critical role to satisfy the stress tensor covariantly conservation condition. Singularity of the Hadamard Green function is assumed to be has logarithmic type same as the Green function in 2d Minkowski flat space time satisfying the general covariance condition. Our procedure has an advantage with respect to other methods such as zeta function regularization: Applying energy conditions (SEC, WEC, NEC) on the renormalized quantum matter dilaton field stress tensor we obtained dynamical equations of the dilaton field  $\Phi$ , quantum vacuum state  $W_0(x)$  and variable cosmological parameter  $\lambda(x)$  respectively. This is still an important problem in the Hadamard renormalization prescription used in general form of background metric in higher than 2 dimensions. In slow varying limits of quantum fields our obtained anomaly trace satisfies the well known one which is obtained from zeta regularization method.

## REFERENCES

1. N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved space*, Cambridge University press, Cambridge, England, (1982).
2. L. Parker and D. Toms *Quantum Field Theory in Curved space-time*, Cambridge University Press, Cambridge (2009).
3. R. M. Wald, Phys. Rev. D17, 1477 (1978).
4. S. M. Christensen and S. A. Fulling, Phys. Rev. D15, 2088 (1977).
5. V. Mukhanov, A. Wipe and A. Zelnikov, Phys. Lett B332, 283 (1994), hep-th/9403018.
6. R. Balbinot and A. Fabbri, Phys. Rev. D59, 044031 (1999), hep-th/9807123.
7. W. Kummer, H. Liebl and D. V. Vassilevich, Mod. Phys. Lett. A12, 2683 (1997), hep-th/9707041.
8. R. Balbinot and A. Fabbri, Phys. Lett. B459, 112 (1999), gr-qc/9904034
9. P. Thomi, B. Isaak, and P. Hajicek, Phys. Rev. D30, 1168 (1984).
10. R. Haag. H. Narnhofer and U. Stein, Commun. Math. Phys. 94, 219, (1984).
11. A. I. Janis, E. T. Newman and J. Winicour, Phys. Rev. Lett. 20, 878, (1968).
12. K. S. Virbhadra, D. Narasimha and S. M. Chitre, Astron. Astrophys. 337, 1-8 (1998).
13. M. Gasperini, *Elements of String Cosmology*, Cambridge University press (2007).
14. M. R. Brown, J. Math. Phys.25(1), 136 (1984).
15. D. Bernard and A. Folacci, Phys. Rev. D34, 2286 (1986).
16. H. Ghafarnejad and H. Salehi, Phys. Rev. D 56, 4633, (1997); 57, 5311 (E) (1998).
17. R. Bousoo and S. W. Hawking, Phys. Rev. D56, 7788 (1997).
18. W. Kummer, H. Liebl and D. V. Vassilevich, Phys. Rev. D58, 108501 (1998), hep-th/9801122.
19. S. Ichinose, Phys. Rev. D57, 6224, (1998), hep-th/9707025.
20. E. Elizalde, S. Naftulim, S. O. Odintsov, Phys. Rev. D49, 2852 (1994), hep-th/9308020.
21. S. Nojiri and S. D. Odintsov, Mod. Phys. Lett. A12, 2083 (1997), hep-th/9706009.
22. S. Nojiri and S. D. Odintsov, Phys. Rev. D57, 2363 (1998), hep-th/9706143.
23. S. Nojiri and S. D. Odintsov, Phys. Lett B416, 85 (1998), hep-th/9708139.
24. S. Nojiri and S. D. Odintsov, Phys. Lett. B426, 29 (1998), hep-th/9801052.
25. S. Nojiri and S. D. Odintsov, Phys. Rev. D57, 4847 (1998), hep-th/9801180.
26. S. J. Gates Jr, T. Kadoyoshi and S. D. Odintsov, Phys. Rev. D58, 084026 (1998), hep-th/9802139.
27. P. Van Nieuwenhuizen, S. Nojiri and S. D. Odintsov, Phys. Rev. D60, 084014 (1999), hep-th/9901119.
28. S. Nojiri and S. D. Odintsov, Int. J. Mod. Phys. A16, 1015 (2001), hep-th/0009202.
29. J. S. Dowker, Class. Quantum Grav 15, 1881 (1998), hep-th/9802029.

# Authors Index

|                  |          |                  |             |                         |              |
|------------------|----------|------------------|-------------|-------------------------|--------------|
| Adamopoulou, E.  | 41       | Kalimulin, I.    | 151         | Redou, P.               | 109          |
| Akgol, O.        | 121, 225 | Karaaslan, M.    | 121, 225    | Remoundou, C.           | 41           |
| Akhunov, R.      | 151      | Kazanskiy, N. L. | 50, 87, 134 | Riaza, R.               | 83           |
| Al-Shalfan, K.   | 232      | Khonina, S. N.   | 50, 87      | Róka, R.                | 58           |
| Amghar, A.       | 147      | Kim, J.          | 130         | Sabah, C.               | 54, 121, 225 |
| Ammar, A.        | 232      | Kim, S.-I.       | 130         | Sahsah, H.              | 147          |
| Assink, N.       | 30       | Kim, Y. H.       | 130         | Salov, V.               | 151          |
| Baik, J. H.      | 75       | Kollmitzer, C.   | 103         | Sandulyak, Al.          | 228          |
| Both, I.         | 45       | Komnatnov, M.    | 151         | Sandulyak, An.          | 228          |
| Carpentieri, B.  | 25       | Kosmides, P.     | 41          | Sandulyak, D.           | 228          |
| Chew, L. Y.      | 196      | Křišťálová, D.   | 116         | Savelyev, D. A.         | 87           |
| Chung, N. N.     | 196      | Kuksenko, S.     | 151         | Schartner, P.           | 103          |
| Ciftci, Y. O.    | 163      | Kutěj, L.        | 116         | Selberherr, S.          | 17           |
| Degtyarev, S. A. | 50       | Lai, C. H.       | 196         | Semina, O.              | 228          |
| Demestichas, K.  | 41       | Le Gal, C.       | 109         | Serafimovich, P. G.     | 134          |
| Demirel, E.      | 225      | Le Yaouanq, S.   | 109         | Sheinerman, A. G.       | 193          |
| Derra, M.        | 147      | Lin, C.-H.       | 220         | Sick, B.                | 21           |
| Dincer, F.       | 121, 225 | Lin, J.-M.       | 220         | Song, H. Y.             | 75           |
| Djebabra, M.     | 215      | Loumiotis, I.    | 41          | Stodola, P.             | 116          |
| Dmitriev, A. L.  | 237      | Makarov, A.      | 17          | Sugiarto, H. S.         | 196          |
| Embrechts, M. J. | 21       | Mallaiah, K.     | 140         | Surovtsev, R.           | 151          |
| Evecen, M.       | 163      | Mazal, J.        | 116         | Sverdlov, V.            | 17           |
| Fenzl, T.        | 103      | Melkozerov, A.   | 151         | Tang, Q.                | 205          |
| Fox, J.-P.       | 30       | Mogulkoc, Y.     | 163         | Theologou, M.           | 41           |
| Ganorkar, S.     | 130      | Mokhnache, L.    | 215         | Tisseau, J.             | 109          |
| Gazizov, A.      | 151      | Munoz, S. S.     | 173         | Ukachukwu, U. C.        | 220          |
| Gazizov, T.      | 151      | Nobile, L.       | 69          | Unal, E.                | 121, 225     |
| Ghaffarnejad, H. | 241      | Nobile, S.       | 69          | Van Der Lubbe, R. H. J. | 30           |
| Ghosh, J.        | 17       | Orlov, P.        | 151         | Wang, Xi.               | 205          |
| Groppen, V. O.   | 210      | Ovid'ko, I. A.   | 193         | Wang, Xu                | 205          |
| Gunduz, O. T.    | 54       | Perkovac, M.     | 92          | Windbacher, T.          | 17           |
| Hedayati, R.     | 186      | Pizzolante, R.   | 25          | Zabolotsky, A.          | 151          |
| Hu, J.           | 205      | Podhorec, M.     | 116         | Zapata, S. C.           | 173          |
| Ivan, A.         | 45       | Porfiriev, A. P. | 50          | Zeghichi, L.            | 215          |
| Jahanbakhshi, M. | 186      | Ramachandram, S. | 140         |                         |              |
| Jakóbczak, D. J. | 124      | Rass, S.         | 103         |                         |              |