International Journal on Advances in Systems and Measurements

2020 vol. 13 nr. 1&2

The *International Journal on Advances in Systems and Measurements* is published by IARIA.
ISSN: 1942-261x
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Systems and Measurements
Volume 13, Number 1 & 2, 2020

Jaime Calvo-Gallego, University of Salamanca, Spain
Maria-Dolores Cano Baños, Universidad Politécnica de Cartagena,Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Vítor Carvalho, Minho University & IPCA, Portugal
Irinela Chilibon, National Institute of Research and Development for Optoelectronics, Romania
Soolyeon Cho, North Carolina State University, USA
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Denis Collange, Orange Labs, France
Noelia Correia, Universidade do Algarve, Portugal
Pierre-Jean Cottinet, INSA de Lyon - LGEF, France
Paulo Estevao Cruvinel, Embrapa Instrumentation Centre - São Carlos, Brazil
Marc Daumas, University of Perpignan, France
Jianguo Ding, University of Luxembourg, Luxembourg
António Dourado, University of Coimbra, Portugal
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France
Matthew Dunlop, Virginia Tech, USA
Mohamed Eltoweissy, Pacific Northwest National Laboratory / Virginia Tech, USA
Paulo Felisberto, LARSyS, University of Algarve, Portugal
Javad Foroughi, University of Wollongong, Australia
Miguel Franklin de Castro, Federal University of Ceará, Brazil
Mounir Gaidi, Centre de Recherches et des Technologies de l'Energie (CRTEn), Tunisie
Eva Gescheidtova, Brno University of Technology, Czech Republic
Tejas R. Gandhi, Virtua Health-Marlton, USA
Teodor Ghetiu, University of York, UK
Franca Giannini, IMATI - Consiglio Nazionale delle Ricerche - Genova, Italy
Gonçalo Gomes, Nokia Siemens Networks, Portugal
Luis Gomes, Universidade Nova Lisboa, Portugal
Antonio Luis Gomes Valente, University of Trás-os-Montes and Alto Douro, Portugal
Diego Gonzalez Aguilera, University of Salamanca - Avila, Spain
Genady Grabarnik,CUNY - New York, USA
Craig Grimes, Nanjing University of Technology, PR China
Stefanos Gritzalis, University of the Aegean, Greece
Richard Gunstone, Bournemouth University, UK
Jianlin Guo, Mitsubishi Electric Research Laboratories, USA
Mohammad Hammoudeh, Manchester Metropolitan University, UK
Petr Hanáček, Brno University of Technology, Czech Republic
Go Hasegawa, Osaka University, Japan
Henning Heuer, Fraunhofer Institut Zerstörungsfreie Prüfverfahren (FhG-IZFP-D), Germany
Paloma R. Horche, Universidad Politécnica de Madrid, Spain
Vincent Huang, Ericsson Research, Sweden
Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek - Hannover, Germany
Travis Humble, Oak Ridge National Laboratory, USA
Florentin Ipate, University of Pitesti, Romania
Imad Jawhar, United Arab Emirates University, UAE
Terje Jensen, Telenor Group Industrial Development, Norway
Liudi Jiang, University of Southampton, UK
Kenneth B. Kent, University of New Brunswick, Canada
Fotis Kerasiotis, University of Patras, Greece
Andrei Khrennikov, Linnaeus University, Sweden
Alexander Klaus, Fraunhofer Institute for Experimental Software Engineering (IESE), Germany
Andrew Kusiak, The University of Iowa, USA
Vladimir Laukhin, Institució Catalana de Recerca i Estudis Avançats (ICREA) / Institut de Ciencia de Materials de Barcelona (ICMAB-CSIC), Spain

Kevin Lee, Murdoch University, Australia
Wolfgang Leister, Norsk Regnesentral (Norwegian Computing Center), Norway
Andreas Löf, University of Waikato, New Zealand
Jerzy P. Lukaszewicz, Nicholas Copernicus University - Torun, Poland
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Stefano Mariani, Politecnico di Milano, Italy
Paulo Martins Pedro, Chaminade University, USA / Unicamp, Brazil
Don McNickle, University of Canterbury, New Zealand
Mahmoud Meribout, The Petroleum Institute - Abu Dhabi, UAE
Luca Mesin, Politecnico di Torino, Italy
Marco Mevius, HTWG Konstanz, Germany
Marek Miskowicz, AGH University of Science and Technology, Poland
Jean-Henry Morin, University of Geneva, Switzerland
Fabrice Mourlin, Paris 12th University, France
Adrian Muscat, University of Malta, Malta
George Oikonomou, University of Bristol, UK
Arnaldo S. R. Oliveira, Universidade de Aveiro-DETI / Instituto de Telecomunicações, Portugal
Aida Omerovic, SINTEF ICT, Norway
Victor Ovchinnikov, Aalto University, Finland
Telhat Özdoğan, Amasya University - Amasya, Turkey
Gurkan Ozhan, Middle East Technical University, Turkey
Constantin Paleologu, University Politehnica of Bucharest, Romania
Matteo G A Paris, Universita` degli Studi di Milano,Italy
Vittorio M.N. Passaro, Politecnico di Bari, Italy
Giuseppe Patanè, CNR-IMATI, Italy
Marek Penhaker, VSB- Technical University of Ostrava, Czech Republic
Juho Perälä, Bitfactor Oy, Finland
Florian Pinel, T.J.Watson Research Center, IBM, USA
Ana-Catalina Plesa, German Aerospace Center, Germany
Miodrag Potkonjak, University of California - Los Angeles, USA
Alessandro Pozzebon, University of Siena, Italy
Vladimir Privman, Clarkson University, USA
Mohammed Rajabali Nejad, Universiteit Twente, the Netherlands
Konandur Rajanna, Indian Institute of Science, India
Nageswara Rao, Oak Ridge National Laboratory, USA
Stefan Rass, Universität Klagenfurt, Austria
Candid Reig, University of Valencia, Spain
Teresa Restivo, University of Porto, Portugal
Leon Reznik, Rochester Institute of Technology, USA
Gerasimos Rigatos, Harper-Adams University College, UK
Luis Roa Oppliger, Universidad de Concepción, Chile
Ivan Rodero, Rutgers University - Piscataway, USA
Lorenzo Rubio Arjona, Universitat Politècnica de València, Spain
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance, Germany
Subhash Saini, NASA, USA
Mikko Sallinen, University of Oulu, Finland
Christian Schanes, Vienna University of Technology, Austria
Rainer Schönbein, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Germany
Cristina Seceleanu, Mälardalen University, Sweden
Guodong Shao, National Institute of Standards and Technology (NIST), USA
Dongwan Shin, New Mexico Tech, USA

Larisa Shwartz, T.J. Watson Research Center, IBM, USA
Simone Silvestri, University of Rome "La Sapienza", Italy
Diglio A. Simoni, RTI International, USA
Radosveta Sokullu, Ege University, Turkey
Junho Song, Sunnybrook Health Science Centre - Toronto, Canada
Leonel Sousa, INESC-ID/IST, TU-Lisbon, Portugal
Arvind K. Srivastav, NanoSonix Inc., USA
Grigore Stamatescu, University Politehnica of Bucharest, Romania
Raluca-Ioana Stefan-van Staden, National Institute of Research for Electrochemistry and Condensed Matter, Romania
Pavel Šteffan, Brno University of Technology, Czech Republic
Chelakara S. Subramanian, Florida Institute of Technology, USA
Sofiene Tahar, Concordia University, Canada
Muhammad Tariq, Waseda University, Japan
Roald Taymanov, D.I.Mendeleyev Institute for Metrology, St.Petersburg, Russia
Francesco Tiezzi, IMT Institute for Advanced Studies Lucca, Italy
Wilfried Uhring, University of Strasbourg // CNRS, France
Guillaume Valadon, French Network and Information and Security Agency, France
Eloisa Vargiu, Barcelona Digital - Barcelona, Spain
Miroslav Velev, Aries Design Automation, USA
Dario Vieira, EFREI, France
Stephen White, University of Huddersfield, UK
Shengnan Wu, American Airlines, USA
Qingsong Xu, University of Macau, Macau, China
Xiaodong Xu, Beijing University of Posts & Telecommunications, China
Ravi M. Yadahalli, PES Institute of Technology and Management, India
Yanyan (Linda) Yang, University of Portsmouth, UK
Shigeru Yamashita, Ritsumeikan University, Japan
Patrick Meumeu Yomsi, INRIA Nancy-Grand Est, France
Alberto Yúfera, Centro Nacional de Microelectronica (CNM-CSIC) - Sevilla, Spain
Sergey Y. Yurish, IFSA, Spain
David Zammit-Mangion, University of Malta, Malta
Guigen Zhang, Clemson University, USA
Weiping Zhang, Shanghai Jiao Tong University, P. R. China

International Journal on Advances in Systems and Measurements
Volume 13, Numbers 1 & 2, 2020

CONTENTS

Development of Front-End Readout Electronics System for the ALICE HMPID and Charged-Particle Veto Detectors

Clive Seguna, Edward Gatt, Ivan Grech, Owen Casha
Department of Microelectronics and Nanoelectronics
University of Malta
Msida, Malta
e-mail: {clive.seguna, edward.gatt, ivan.grech, owen.casha}@um.edu.mt

Giacinto De Cataldo
Department of Physics
University of Bari
Bari Italy
e-mail:Giacinto.de.Cataldo@cern.ch

Yuri Kharlov, Artem Shangaraev
Department of Physics
Institute for High Energy Physics, Protvino 142281, Russia
e-mail: {Yuri.Kharlov, Artem.Shangaraev}@cern.ch

***Abstract*— Luminosity of lead-ion collisions at the Large Hadron Collider will be increased in the forthcoming Run 3 to 6×10^{27} cm^2s^{-1}, corresponding to an average inelastic interaction rate of 50 kHz. At the same time, paradigm of data taking of the ALICE experiment changes aiming to collect and process all interaction data, which represents an increase in data sample rate by two orders of magnitude with respect to the present system. This requirement demands a reliable readout electronic system with an increase in data bandwidth, strict timing constraints, and low power consumption. This work presents the hardware architecture of a newly developed front-end readout electronic system for the Charged-Particle Veto detector, located in the Photo Spectrometer at the A Large Ion Collider Experiment situated at the largest European facility for Nuclear Research, CERN. The developed front-end hardware architecture enables the simultaneous readout of 23,040 cathode pad channels for amplitude analysis, contributing a ten-fold increase in bandwidth when compared to prior system. Main contributions to this achievement include the re-design of highly dense interconnect printed circuit boards, use of 3.125 Gbps data links and the implementation of a radiation tolerant firmware architecture using low power 28 nm field programmable gate arrays. Measurement results indicate that the newly developed data acquisition electronic system satisfies the target detector readout rate requirements. This paper discusses the firmware and hardware implementation details, followed by the presentation of the performance measurement results for the recently developed Charged-Particle Veto detector front-end readout topology when compared to other particle detector electronic systems.**

Keywords- electronics; readout; detector; FPGA; VHDL.

I. INTRODUCTION

This journal paper is an extension of the work originally presented at 11th Conference on Advances in Circuits, Electronics and Micro-electronics CENICS 2018 [1]. This Section explains the overall system architecture and performance of the prior CPV readout electronic system operated in the ALICE experiment during the LHC Run 1 and Run 2 (2010-2018). Particle detectors such as Charged Particle Veto Detector (CPV) and High Momentum Particle Identification (HMPID) are required to find and possibly identify all the particles emerging from a scattering event. The design of such detectors for colliding beam experiments presents several challenges. These include minimization of materials because of limited space and services, low-power consumption and potentially high data rates and reliability when running a specific radiation tolerance level. Particle detectors rely on custom designed electronic hardware for identifying specific types of particles. Although detectors appear to be very different, basic principles of the readout apply to all. They consist of a signal current sensor whose output after integration yields to a charge proportional to the amount of energy collected.

Every particle detector needs to a have a custom designed Front-End electronics (FEE) system to be able to detect specific beam types and particle characteristics, therefore, the equipment must be modular, and adaptable. Additionally, the design criteria for a particle detector depends on application, energy resolution, rate or timing capability requirements, and sensing positioning. Large-scale systems additionally impose compromises on power consumption, scalability and easy monitoring with a reduction in maintenance cost. Today, multi-channel systems are additionally required in many fields. In large systems power dissipation and size are critical, so systems are not necessarily designed for optimum noise, but adequate noise, and circuitry needs to be tailored to specific detector requirements. The present CPV detector FEE limits the present readout rate requirements and luminosity levels, therefore, this results in the need of developing a newly customized FEE system that meets specific design criteria.

The CPV is a Multi-Wire Proportional Chamber (MWPC) with cathode-pad readout located in the A Large Ion Collider Experiment (ALICE) [2]. It is used to suppress detection of charged particles hitting the front surface of the Photon Spectrometer (PHOS) in order to improve photon identification [3]. CPV charged-particle detection efficiency is better than 99%. The spatial resolution of the reconstructed impact point is about 1.54 mm along the beam direction and 1.38 mm across the beam. The CPV pad electronics is identical to the one used for the A Large Ion Collider Experiment (ALICE) High Momentum Particle Identification (HMPID). A primary consideration for PHOS is that it has an optimal performance for measuring photons

in the energy range from few hundred MeV to 100 GeV. The PHOS contains 12,544 detection channels based on lead tungstate crystals. During the LHC Run 2, CPV consisted of one module installed on the top of one of the PHOS module. The CPV module consists of:

- 16 columns;
- 10 cards per column, where each card consists of three ASIC charge amplifier signal condition chips called Gassiplex07-3;
- 48 cathode readout pads per Gassiplex07-3 card;
- 7680 channels of amplitude analysis per module.

The ALICE experiment is dedicated for studying properties of strongly interacting matter created in high-energy heavy ion and proton collisions. The current system still leaves open physics questions that need to be addressed, and these questions relate to, among others, hadronization, nuclei, long range capability correlations and small x-proton structure [4]. CPV electronics consists of dedicated Application Specific Integrated Circuit (ASIC) devices in each column, Gassiplex for analogue signal processing and DiLogic for handling the digitized information. Every column consists of 10 cards with Gassiplex chips, called 3-GAS cards interfaced directly on the backside of the MWPC cathode. A customized electronic board called DiLogic contains five channels of 12-bit Analogue-to-Digital Converter (ADC) modules and five DiLogic (5-DIL) processors [5].

Each column contains 480 pads connected with two 5-DIL cards and a group of Field-Programmable Gate Arrays (FPGAs) called column and segment controllers for processing various control signals, and additionally provide the necessary interface to the Data Acquisition (DAQ) module and Central Trigger Processor (CTP). Further, CTP is responsible for the generation of three trigger level signals L0, L1 and L2. At power-on an order of 1000 events are collected with the zero-suppression turned off so to measure the pedestal levels. For each channel, the thresholds are calculated as Thr(j) = Ped(j) + N(j)*Sig(j), where Ped(j) is the pedestal mean for channel j, Sig(j) is the corresponding r.m.s. value and N(j) is a parameter, usually set to 3. The pedestals and thresholds tables are downloaded into the DiLogic chip memory and finally the zero-suppression is turned on. The system is then ready for normal DAQ operation. The L0 trigger is used to store the analogue information inside the Gassiplex chip and start the multiplexing sequence. Then the pulse height information for analogue channels above threshold is acquired after pedestal subtraction, and then stored in the DiLogic internal First-In-First-Out (FIFO) memory (512 x 18-bit words) together with the corresponding analogue channel address. The data from two 5-DIL cards each having 5 digital signal processor chips connected in a daisy chain are collected in the column FIFO as shown in Figure 1. Finally, at arrival of the L2 trigger signal event data is transferred to the DAQ experiment through a 2.125 Gbps optical Detector Data Link (DDL) for further processing.

A typical event size consists of 1.3 Kbytes for Pb-Pb particles. With just firmware upgrade the maximum possible event readout rate that the present detector electronics can reach is 10 kHz for an occupancy of 1%, therefore, due to this technical limitation, a new front-end readout electronic system has been developed to collect more than 10 nb^{-1} of Pb-Pb collisions at luminosities of up to 6 x 10^{27} cm^{-2} s^{-1}. As shown in Figure 2 the busy time is the period, where the FPGA busy signal goes high, from the arrival of the L0 trigger to the end of the transmission of an event. Therefore, it includes the waiting time of a L2 trigger (about 108 µs) and the transmission time depending on the number of words from the 32-bit data bus fbD[31..0]. The fixed part consists of headers and markers:

- Common Data Header (CDH) (10 words, 40 bytes);
- CPV header (5 words, 20 bytes);
- Column headers (24 words, 96 bytes);
- DILOGIC markers (240 words, 960 bytes);
- Segment markers (3 words, 12 bytes).

The data block transmitted by the frontend readout electronics to DAQ system consists of a fixed part and another of variable length.

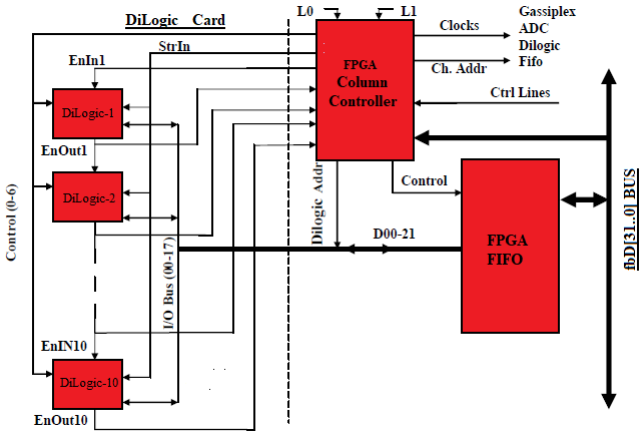


Figure 1. Block Diagram of Column Controller.

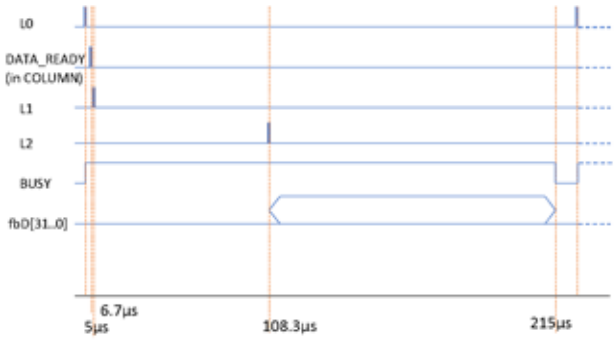


Figure 2. Timing diagram for the acquisition of a physics event-data, showing the relationship between various detector data and control signals.

Therefore, each event consists of total of 1128 bytes, currently being transferred at a rate of not more than 5 kHz by readout electronics. The variable event rate depends on the number of activated channels.

The plot shown in Figure 3 illustrates that the detector event readout rate decreases with the number of transmitted or increased detector occupancy. On the prior system, the Frontend Electronics (FEEs) implements a full-duplex Serial Interface Unit (SIU). The main SIU task is to transmit event data, receive control commands from Read-out Receiver Card (RORC) or Destination Interface Unit (DIU) located on DAQ via DDL.

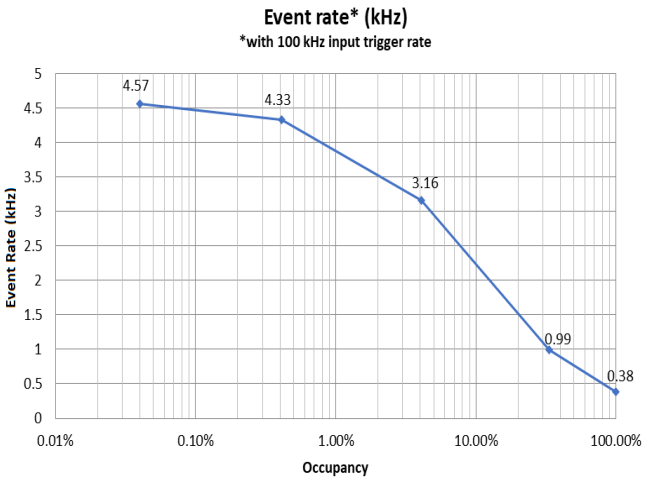


Figure 3. Prior System Readout rate versus Occupancy with 100 kHz trigger rate [6].

The 2.125 Gbps DDL optical interface has two interfaces: the FEE-SIU interface and the RORC - DIU DDL interface. Data transfer from DIU to computer farm for further processing is done through PCI-X communication interface as shown in Figure 4.

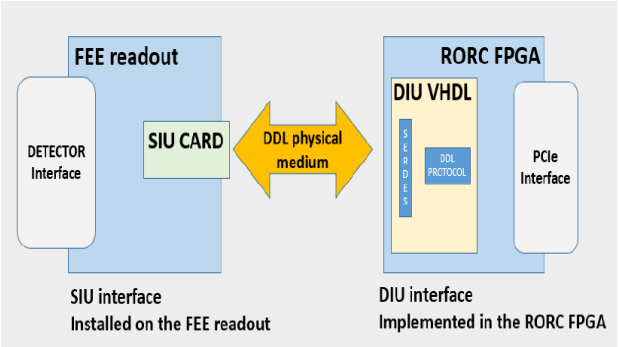


Figure 4. DDL components in data taking configuration between FEE and DAQ [7].

The present CPV readout architecture contributes to a low particle readout rate of 5 kHz or 200 μs busy time. The ALICE experiment is scheduled to start running with an interaction rate of 50 kHz of all Pb-Pb events in 2021 (Run3). Therefore, the installation of the newly produced CPV electronic system, which includes 3 modules, each having 8 optimized FPGA column controller cards shall contribute to an increase in bandwidth, data rate and drastic reduction in busy time by ten-fold. In Run3, CPV will consist of 3 modules installed in front of 3 PHOS modules at sectors 260-280°, 280-300°, 300-320°. The main goal of CPV is to improve photon identification in PHOS via charged cluster rejection, therefore, events taken by PHOS and CPV should be in complete synchronization. As PHOS plans to upgrade readout rate up to 40-50 kHz, CPV should not be slower than PHOS but possibly faster.

Current readout rate of the old CPV electronics can be increased from 5 kHz to 10 kHz just by a firmware upgrade of the readout board. Further increase of readout rate is not possible without hardware upgrade. Therefore, new readout cards had to be designed and produced keeping the same data flow as the one implemented in the previous system but taking a chance to speed up by tenfold the read-out event rate.

The rest of the paper is structured as follows. Section II gives an overview of the newly developed readout electronic hardware including two types of printed circuit boards: FPGA based Column controller and passive motherboard called Segment. Section III provides a description of the implemented and optimized column controller electronic hardware. Simulation and measurement results are shown in the following sections. Finally, Section X presents the novelty of this work, other conclusions and future work.

II. New System Architecture

The evaluation of various FPGA-based electronic boards that are currently available in the market was performed, and it was concluded that no FPGA electronic card with the required features is available for the development of this new CPV detector readout electronics. Therefore, a complete re-design and implementation of all the detector electronic controller cards had to be completed.

As requested by ALICE collaboration, the objective is to preserve the charge-sensitive amplifiers (3-GAS cards) and Analogue-to-Digital Converter (ADC) together with the 5-DIL processors, 96 cards in total, while upgrading the column controllers (CCs) (from16 CCs/module to 8 CCs/module).

Every column controller shall simultaneously process two columns of 480 3-Gassiplex channels. In total leading to 24 FPGA column controller cards for all the three CPV modules, with a total of 7680 3-Gassiplex channels per module. Additionally, segment motherboards shall be upgraded to eight per module leading to a total of 24 segment boards.

The upgrade for the new CPV readout electronic system includes the parallel readout of all column controllers via 3.125 Gbps FPGA unidirectional transceiver serial links and integration with the ALICE Inner Tracking System Readout Unit (ITS-RU) [8]. The ITS-RU frontend electronic system is divided into a modular Readout Units (RU), identical for each layer. Each Readout unit controls an entire stave, including power to the sensors (through custom-made power units) [9].

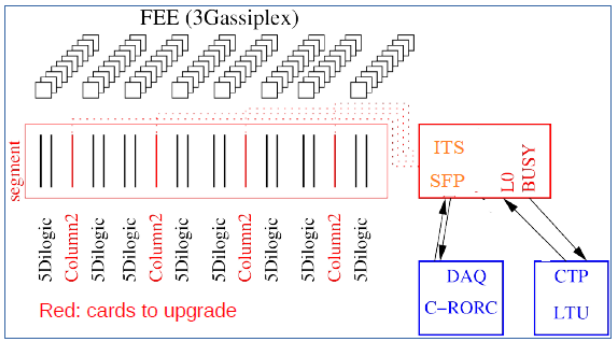


Figure 5. Block diagram for the upgrade of CPV and HMPID electronics.

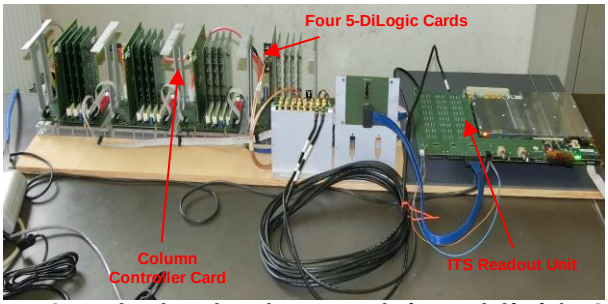


Figure 6. Developed Readout electronic cards for one half of the CPV module (4 Segments, 4 CCs, 16 5-DIL cards and 1 ITS-RU).

Figure 7. CPV module for the simultaneous Readout of 7,680 analogue channels (8 CCs, 32 5-DIL cards and 8 Transceiver links).

The ITS-RU shown in Figure 8 will serve to simultaneously transmit event data from all 24 CPV/HMPID column controllers to the Online computing system (O^2) using the SMA-based connectors Gigabit transceivers. All upgraded cards are shown in Figure 5 and Figure 6.

III. Optimzed FPGA Column Controller

The layout of the optimized FPGA CPV column controller card is shown in Figure 9 and Figure 10. The card has a 364 pins High-Speed Mezzanine edge connector, includes a Cyclone V GX Intel FPGA, three power supply voltage regulators of 3.3 V, 1.1 V and 2.5 V, a Low-Voltage Differential Signalling (LVDS) Fire-Fly connector, and high-speed full-duplex transceiver links for command and event data transfer between FPGA column controllers and ITS-RU. The newly adopted architecture allows the simultaneous readout of two columns, where each column contains 480 Gassiplex channels, an improvement of two-fold when compared with the prior system. The two main simultaneous operations that are required to be performed by the newly developed FPGA-based controller card include frontend and backend operations. The various modes of operations are listed in Table I. The firmware and hardware of the CC allow to work with four 5-DIL cards in parallel.

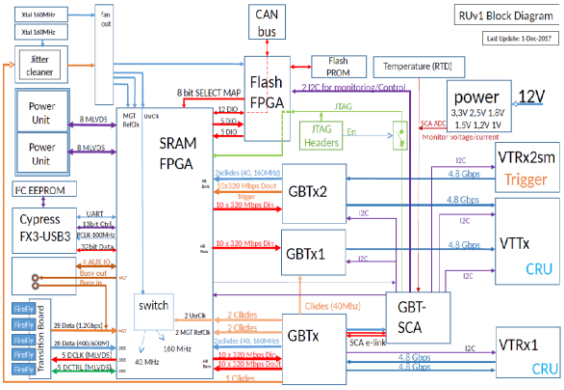


Figure 8. Block Diagram for ITS-RU.

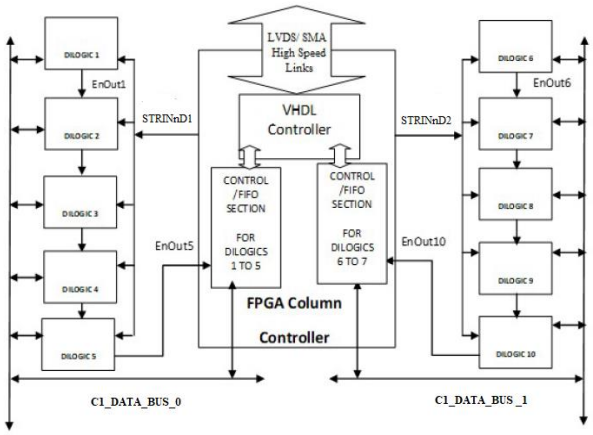


Figure 9. Block Diagram for FPGA-based controller card [10].

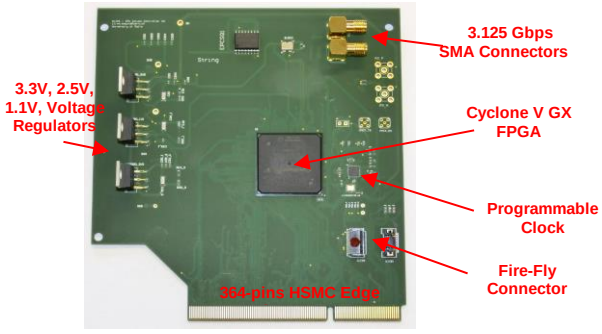


Figure 10. Manufactured FPGA column controller card for CPV Readout electronics.

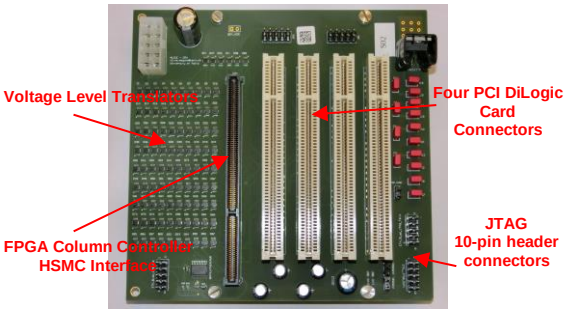


Figure 11. Manufactured Segment card for CPV Readout electronics.

Figure 11 illustrates the main interfaces for the developed segment card, for interfacing column controller and four 5-DIL processor cards, as shown in Figure 9.

TABLE I. FUNCTION CODES FOR FPGA CONTROLLER CARD.

Operating Modes/Codes	Description	Type of Operation
"1010"	Analogue Readout	Back-End
"0111"	Data Acquisition	Front-end
"1000"	DiLogic Pattern Read Out	Back-End
"1110"	DiLogic Write Configuration	Back-End
"1111"	DiLogic Read Configuration	Back-End
"0000"	Test Mode - DiLogic	Front-End

IV. COLUMN CONTROLLER CARD - FRONT-END OPERATIONS

A. *Data Acquisiton and Test mode*

As shown in Figure 12, prior starting a data acquisition the DiLogic processor should have the pedestal threshold memory filled with pedestal and operating threshold values for each of the 480 channels.

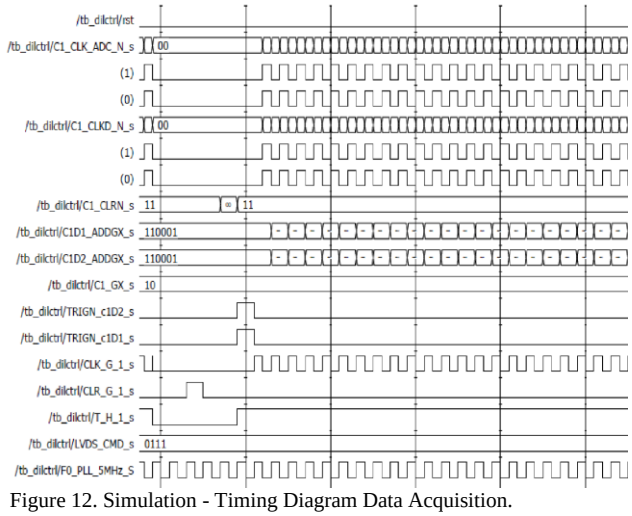


Figure 12. Simulation - Timing Diagram Data Acquisition.

The frontend card must activate the CLR_N and RST for initializing DiLogic processor prior data taking. Every event starts with a pulse on TRIGN pin and several clock cycles on the C1_CLKD_N and CLK_ADC_N pins depending on the number of channels indicated by the binary value "10" on the C1_GX pin. An extra clock cycle is necessary to store the end-event word and turn off the readout of an event. Timing diagram for the complete data acquisition sequence is shown in Figure 12.

Similar timings can be applied for Test-mode operation but instead of asserting the amplitude and channel address on the input pins, they should be filled through the I/O Digital bus, respectively with channel address on C1_DATA_BUS0/1 (17:12) on D17-D12 and Amplitude on C1_DATA_BUS0/1 (11:0).

V. COLUMN CONTROLLER CARD - BACK-END OPERATIONS

A. *DiLogic Read and Write Configurations*

The DiLogic threshold and offset memory can be read using the function code "1111", where pins ENIN1N, ENIN5N are set to low and STRINnD1, STRINnD2 strobe cycles are applied.

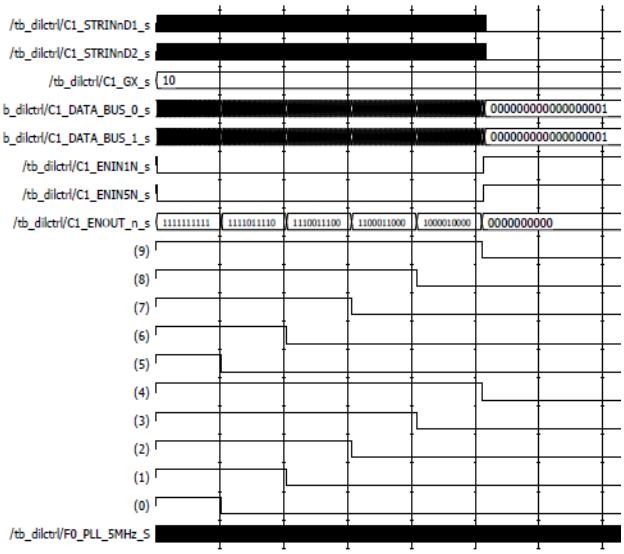


Figure 13. Simulation - Timing Diagram for DiLogic Read Configuration.

The data will appear on the data bus after the falling edge of strobe and will stay stable until the rising edge of STRINnD1, and STRINnD2 pins as shown in Figure 13. The threshold and offset memory of DiLogic chip can be loaded using the function code "1110" shown in Figure 14, where the ENIN1N and ENIN5N pins are set to low and STRINnD1, STRINnD2 strobe cycles are applied.

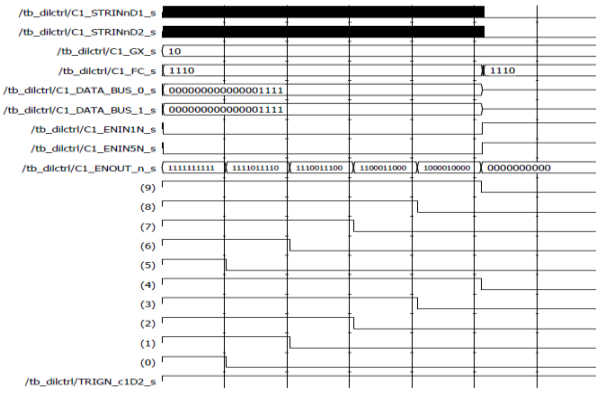


Figure 14. Simulation - Timing Diagram for DiLogic Write Configuration.

The data should be stable on the data bus at the rising edge of STRINnD1 and STRINnD2 signals. A reset daisy chain must be applied at the end of the operation.

B. *DiLogic Analogue Readout*

The DiLogic processor is configured in analogue readout mode using function code "1010" and setting ENIN1N and ENIN5N pins low as shown in Figure 15.

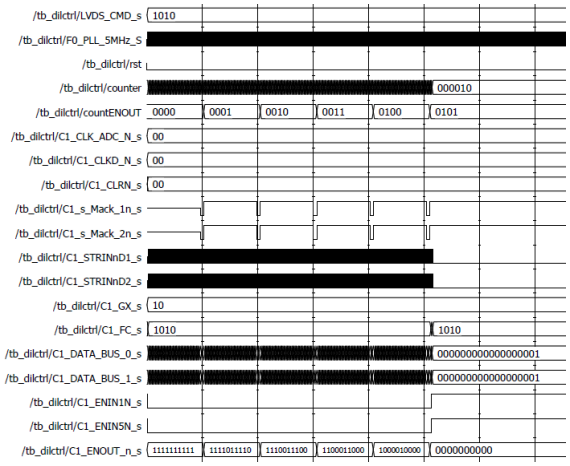


Figure 15. Simulation - Timing Diagram Analogue Readout.

Successive STRINnD1, and STRINnD2 cycles will cause all the daisy chained DiLogic chips to place their digitised data on the data bus one at a time, starting with the first module in the chain.

C. *DiLogic Pattern Readout*

To perform the pattern readout of DiLogic chip Bit-Map memory, the operation code must be set to "1000" and both ENIN1N, ENIN5N pins must be low. While STRINnD1, and STRINnD2 are set to low, the patterns will appear on the data bus. The readout sequence will be the same as the analogue readout, it will be finished when the last module drives its ENOUT_N pin low. A reset daisy chain must be applied to turn off the ENOUT_N pins. As in the analogue readout mode a pattern delete can be performed.

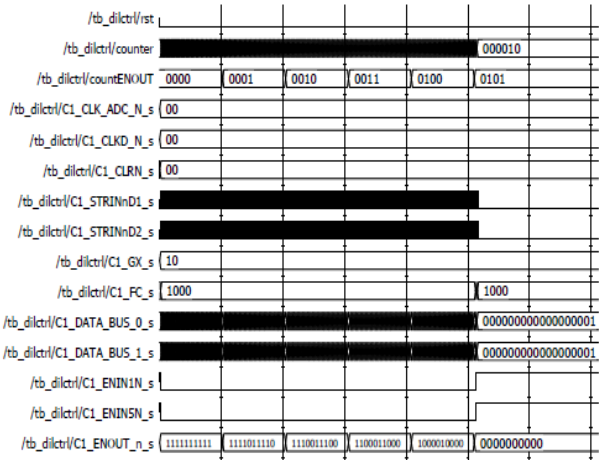


Figure 16. Simulation - Timing Diagram Pattern Readout.

An enable signal is passed from the ENOUT_N pin to the ENIN_N pin of the next chip when the module has finished after transferring its Analogue data of one event on the C1_DATA_BUS (17:0) data bus pins. The MACK_1N and MACK_2N pins indicate the occurrence of the end-event word and the end of the analogue readout on that DiLogic chip.

VI. TRANSCEIVER MEASUREMENT RESULTS

The eye and composite jitter diagram measurement results for the 3.125 Gbps transceiver links are shown in Figures 17 and 18. Measurements were taken using Tektronix 6 GHz real-time scope. The implementation of a 16-bit PRBS generation logic has been applied for validating the physical quality of the high-speed transceiver printed-circuit board links. The implementation of this PRBS generator is based on the linear feedback shift register with the logic XOR and logic AND operations that produces a predefined sequence of 1's and 0's. The measured eye diagram is a common indicator of the quality of signals in high-speed digital transmissions. The test was executed for a duration of 72 hours using 5 m SMA cables. The measured differential input jitter, PRBS pattern at zero crossing is +/- 0.25 UI or 0.5 UI. Rise and Fall times of around 100 ps were measured, with a peak-peak jitter value of +/- 20 ps for a data rate of 3.125 Gbps. The measured Bit-error rate (BER) from the illustrated Bath-Tub plot is 1 x 10^{-12}, which is also an acceptable measurement value according to the digital recommendations standards. Gigabit receivers (3.125 Gbps) are AC coupled with OCT, and use 8b/10b encoder/decoder, byte ordering, and an automatic synchronization state machine. To avoid common-mode noise being generated from the 5m non-

identical differential pair cable properties (i.e., unequal length, diameters, twisting or material) as a remedy for common mode signal, A.C. coupling is used, while for intra-pair skew, can be adjusted through the transceiver slew-rate programmer. AC-coupling allows transceivers to operate with large common-mode offsets.

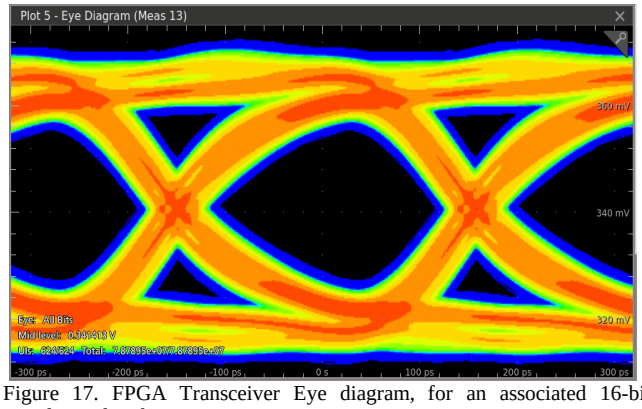


Figure 17. FPGA Transceiver Eye diagram, for an associated 16-bit pseudo-random bit generator pattern.

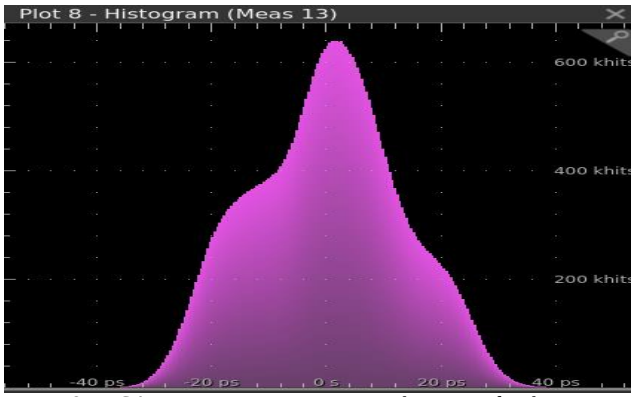


Figure 18. FPGA Transceiver composite jitter histogram for the associated 16-bit pseudo-random bit generator pattern.

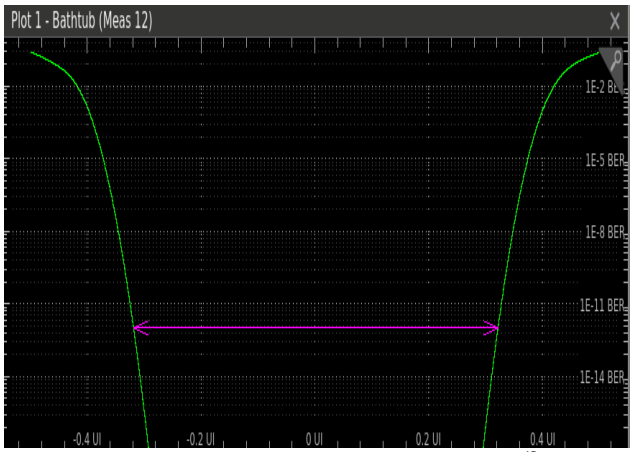


Figure 19. FPGA Transceiver, Bathtub plot showing a 1 x 10^{-12} BER.

The 200 Mbps speed LVDS links are DC coupled with 100-ohm termination resistors connected across the link pairs and placed as close as possible to the receiver. Since for LVDS we are using DC instead of AC coupling then the 8b/10b encoding scheme was not required. The LVDS peak-peak jitter value of +/- 4 ps for a data rate of 200 Mbps.

VII. Readout Chain And Timing Verification

Gassiplex is designed to be connected to wire chambers as well as to silicon strip detectors. The 16-channel Gassiplex chip is an ungated asynchronous device composed of a charge sensitive amplifier (CSA) and signal-conditioning circuitry with a track and hold (T/H) input signal used to store the charges in sample-hold capacitors. Additionally, a burst of clock pulses is required by an external controller to operate the multiplexed readout of the stored charges on a single output line SWAN-OUT. The clear CLR-SROUT pulse is needed to restore the initial state of the sample switches as shown in Figure 20.

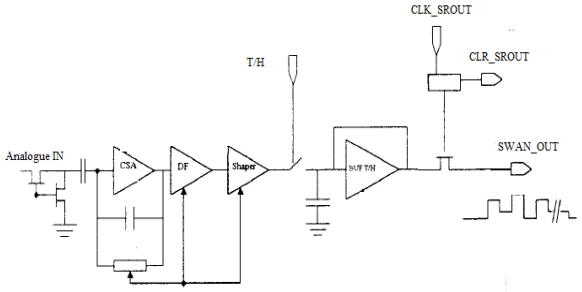


Figure 20. 3-Gassiplex Architecture.

A T/H signal is generated by the FPGA on column controller card as soon as it receives the L0 trigger. When the T/H signal is active the Gassiplex reads the amplitude of the analogue input waveform and holds the value on the T/H buffer capacitor until it is read through a multiplexer. The FPGA provides as well a bunch of clock pulses to read the analogue value and convert it by ADC. The timing of analogue output and clock pulses can be observed from Figure 21. Thus, the maximum amplitude in the hit-channel or pedestal value is then stored in DiLogic processor memory.

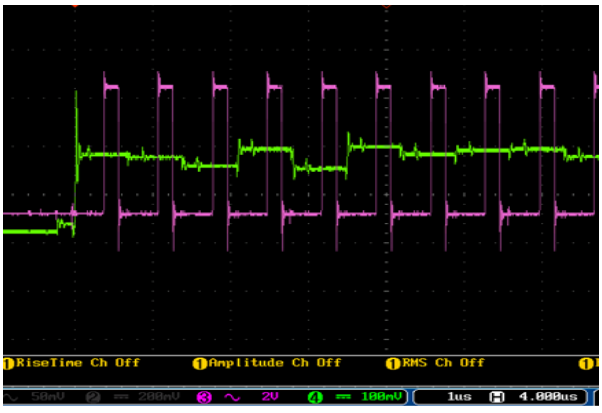


Figure 21. Readout timing specification for 3Gassiplex chip.

The verification of the actual timing measurements is performed using Quartus Embedded Signal Tap Logic

Analyzer synchronously through a global reference clock of 156.25 MHz. Various FPGA registers and nodes were captured and logged on a workstation terminal through the Joint Test Action Group (JTAG) protocol using a 1 MHz clock signal. Figure 22 illustrates the DiLogic chip being put in the analogue readout mode with function code set to "1010" and EnIn_N is low. A burst of Successive StrIn_N cycles will cause all five DiLogic modules in the chain to place their digitised data on the data bus sequentially one at a time. The EnOut_N pin from each DiLogic chip, which is connected to the EnIn_N pin of the next chip is activated when the module has ended transferring its analogue data of one event on the data bus. The Mack_N pin of each DiLogic card indicates the occurrence of the end event word on the 18-bits data bus so to indicate the end of the analogue readout on that DiLogic chip, and immediately start the readout of the next chip. The 18-bits of the end-event word contains the contents of 2 counters: 7-bits from D0 to D6 representing the number of channels above threshold and 11-bits from D7 to D17 indicate the event numbering. The end-event word is indicated in time by the Mack-N output, with the EnOut_N of the last chip going low indicating the termination of the analogue readout operation. All the 5-DiLogic chips must be reset by applying the reset daisy chain code "1101" and an extra Strin_N strobe. The timing diagram for the transfer of event-data between column controller card and ITS-RU is illustrated in Figure 23.

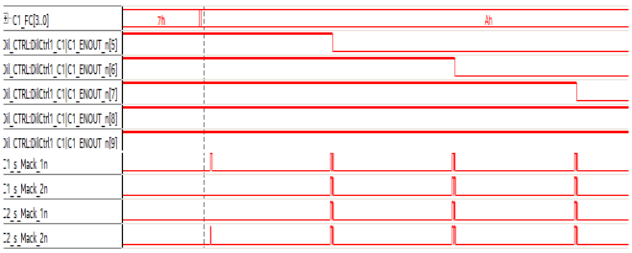


Figure 22. Analogue Readout Timing diagram.

The column controller ACK_Col and rd_req control signals remains high during event data transfer between column controller card and ITS-RU. The end of data transfer is indicated by a pulse being issued on control signal rdyXCVR.

VIII. COLUMN CONTROLLER FIRMWARE DESIGN AND DEVELOPMENT

System firmware consists of a VHDL (VHSIC Hardware Description Language) top-entity FPGA controller module per four 5-DiLogic cards, or two columns. Each FPGA controller consists of four sub-entities DilCtrlC1D1, DilCtrlC1D2, DilCtrlC2D1, DilCtrlC2D2 used for simultaneously controlling all 5-DiLogic cards. Further each DilCtrl controller module implements the control logic for the Gassiplex chips. The Track/Hold (T/H) signal is used to store charges in Gassiplex sampling capacitors using T/H switches.

A burst of clock pulses triggered by the column controller FPGA device is then generated to operate the multiplexed readout of the stored charges on a single output line.

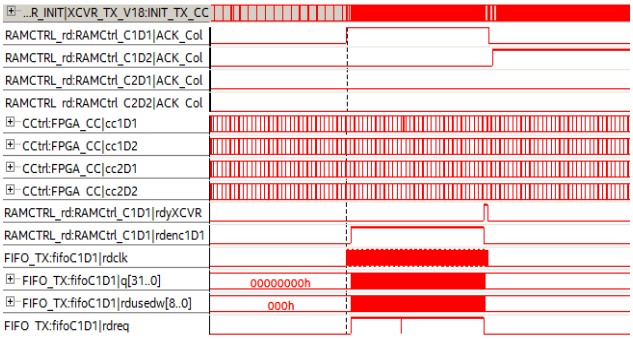


Figure 23. Timing diagram for transfer of event-data between column controller card and ITS-RU.

The ENCol control signal is set high to initiate reading and transfer of event data from the 5-DiLogic cards. Clock frequency for each DilCtrl module is set 10 MHz while that for transceiver is 156.25 MHz. Therefore, components ColC1D1Ctrl, ColC1D2Ctrl, ColC2D1Ctrl, ColC2D2Ctrl implement the synchronization logic and data buffering using a 2kB First-In-First-Out (FIFO) data structure between DilCtrl and 3.125 Gbps transceivers. Additionally, a 200 kbps LVDS link is used to receive L0 CTP command word from ITS-RU unit via the Timing, Trigger Control system (TTC) and issues a Busy flag for the reduction of the overall dataflow.

The Transceiver module performs serialization and de-serializes of event data. The Busy flag is issued from the arrival of the L0 trigger to the end of the transmission of event data. Figure 24 illustrates the complete state-machine for the FPGA controller module. Synchronization between transceiver and FIFO memory is done via flags RDYC1D1, RDYC1D2, RDYC2D1, RDYC2D2, and ACKC1D1, ACKC2D1, ACKC2D2. A high level on SCN_RDY control line indicates the completion of event data transfer to ITS-RU from FIFO memory.

Each event-word contains the selected channel address and digitised amplitude information that need to be transferred via the FPGA transceivers at a rate of 3.125 Gbps then finally to the ITS-RU module for further formatting and transfer to DAQ. DilCtrl Controller activates the respective EnCol pin to initiate data transfer for various functional modes and write to FIFO memory buffer.

The ACK signals are set to '1' to indicate the on-going progress of event-data transfer between FIFO memory and transceiver modules. When reading of data from memory is complete, then RDYx pin is set high by and ACK low. The SCN_RDY control signal indicates that all FIFO memory has been sequentially scanned, and therefore, setting back FPGA controller to IDLE state, waiting for the next event to be transferred to ITS-RU and DAQ.

IX. Preliminary Measurement Results

The busy time of the data collection is mainly defined by the CTP waiting time for the completion of the readout electronics to transmit event data from FEE to DAQ server. The detector busy time due to readout in general depends on the event size.

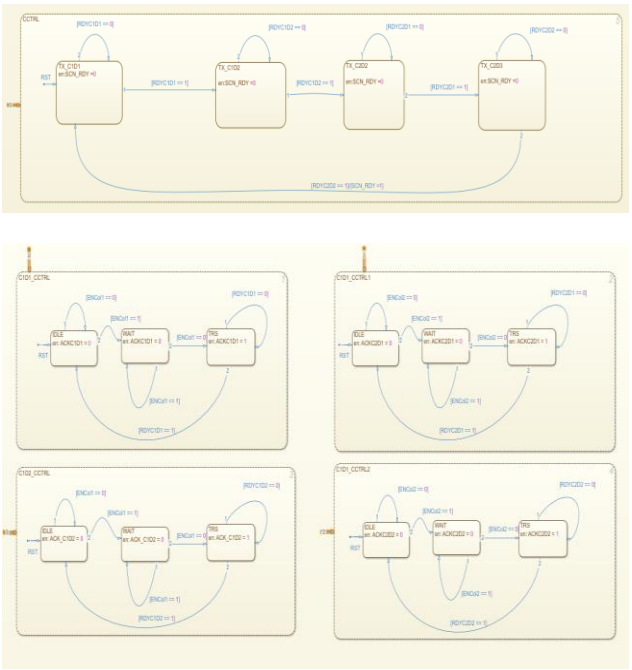


Figure 24. State-Machine for Timing diagram FPGA controller Top-entity.

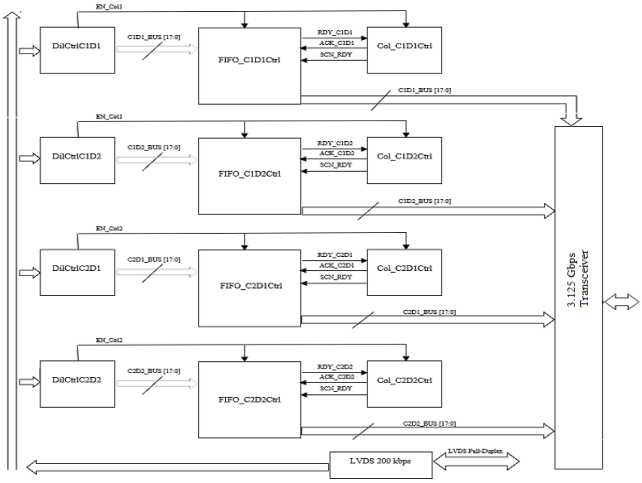


Figure 25. System firmware for FPGA column controller.

With the current firmware and the detector occupancy of 100% the estimated event size from one column controller is 3.8 Kbytes. The corresponding busy time is 33 µs as shown in Table II, which allows to read data at the trigger rate up to 30 kHz. However, the target detector occupancy is 1%, leading to at least a two-fold increase in readout rate, of up-to 70 kHz with enough margin above target requirements. This measurement result is above the required target for a detector occupancy of 1.2 Kbyte Pb-Pb collisions. The maximum event readout rate measurement of the prior system in Run 2 is estimated to be 5 kHz ten-fold slower than this work, to be increased up to 10 kHz just by firmware upgrade. The major contribution to such an improvement is due to the complete re-design of the new electronics hardware architecture leading to the parallel readout of all column controllers, including concurrent readout of 5-DiLogic cards and use of high speed 3.125 Gbps FPGA transceiver links. The location of the proposed new readout electronics presented in this work will be in the ALICE detector where the measured radiation doses are estimated to be 0.1 kRad and 1.9×10^{10} charged particles/cm^2, which puts CPV electronics in a safe operating side by 3 to 4 orders of magnitude [11]. As described in [12], to detect and protect the system against errors caused by SEU in the FPGA memory cells, a threefold way is to be adopted:

- an efficient error detection scheme based on parity check logic;
- 8/10 bits of data coding as part of the transceiver low level protocols;
- a Cyclic Redundancy Check (CRC) will be accompanying data on its way between FEE and ITS-RU board.

The obtained preliminary measurement results shown in Table III indicate an event readout time of ~20 µs (50 kHz) for a detector occupancy of 55% as expected in Run3. Therefore, the newly developed CPV readout electronics contributes to a performance improvement in data transfer rate between column controllers and DAQ by almost a factor of two when compared with the present Scalable Readout Unit (SRU) (~21µs), Time Projection Chamber (TPC), 100µs for High Momentum Particle Identification (HMPID) readout detector electronics as reported in [13], [14], and [15], respectively.

Table II. Measured Elapsed Time For An Event Size of 3.8 Kbytes

Total Busy Time	Elapsed time for Readout of 5-DiLogic Card @10MHz	Elapsed time for Readout of Column Controller to ITS-RU @156.25MHz	Detector Occupancy
33µs	23µs	10 µs	100%

Table III. Readout Rate Comparison With Other Similar Work.

Detector	Estimated Readout Rate (µs)
(this work)	20
SRU [13]	21
TPC [14]	33
HMPID [15]	100

X. Novelty Contribution And Future Work

This paper presented the design of a new CPV Front-end Readout electronics system, which attains the ALICE Readout rate goal of 50 kHz. The preliminary prototype measurements indicate an estimated event Readout rate of at least 50 kHz, as per target value requirements for an occupancy of 1%.

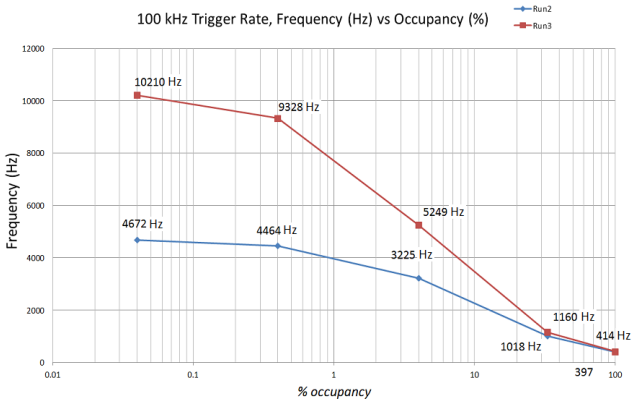


Figure 26. Estimated event readout rate (Hz) for prior System.

The newly designed upgrade offers significantly improved electronics performance. Such an improvement in event readout rate when compared with the prior CPV, TPC, HMPID and SRU readout detector electronics is mainly due to the parallel readout and processing of column controllers and the adopted high-speed transceiver link speeds between DAQ and readout electronics of around 3.125 Gbps. Additionally, the integrated CRC hard Intellectual Property (IP) FPGA block, shall detect and correct errors due to SEU, thus ensuring a reliable operation of the newly developed CPV electronics. A further study to be considered is the evaluation of data reliability versus the improvement in readout trigger rates. Additionally, further reduction in power consumption and area space requirements will be done through the integration of DIL-5 and FPGA column controller functionality into same ASIC device.

After prototype testing and improving in the year of 2018, the full set of new Column controllers and Segment boards was produced and installed on three CPV modules. So, the detector is now ready to work at 50 kHz trigger rate during Run 3. The next step is to replace the old 5-DiLogic cards with 700 nm technology with an ASIC chip for a better system performance, throughput and maintainability.

Acknowledgment

The authors gratefully acknowledge the support of the project under the Tertiary Education Scholarship Scheme (TESS) and the Malta College of Arts, Science, and Technology (MCAST).

References

[1] C. Seguna, E. Gatt, G. Cataldo De, I. Grech and O. Casha, "A New Front-End Readout Electronics for the ALICE Charged-Particle Veto Detector" The Eleventh International Conference on Advances in Circuits, Electronics and Micro-electronics (CENICS 2018) IARIA, Sep. 2018, pp. 11-15, ISSN 2308-426X, ISBN:978-1-61208-664-4

[2] S. Evdokimov et al., “The ALICE CPV Detector”, KEn, vol. 3, no. 1, pp. 260–267, Apr. 2018.

[3] ALICE Collaboration, Technical Design Report of the Photon Spectrometer (PHOS). CERN/LHCC, 1999.

[4] P. Riedler,"Upgrade of the ALICE Detector," in 2nd International Conference on Technology and Instrumentation in Particle Physics (TIPP), pp. 164-169, June 2011, doi: 10.1016/j.phpro.2012.03.707.

[5] J. C. Santiard, “The ALICE HMPID on-detector front-end end readout electronics,” *Nucl. Instrum.Meth.* vol. A518, pp. 498-500, April 2014.

[6] F. Carena, “DDL, the ALICE data transmission protocol and its evolution from 2 to 6 Gb/s,” *JINST*, vol. 10, pp. 2-6, April 2015, doi: 10.1088/1748-0221/10/04/c04008.

[7] J. C. Santiard, K. Maret, "The Gassiplex07-2 integrated front-end analog processor for the HMPID and Dimuon spectrometer of ALICE” The Sixth Workshop on Electronics for LHC Experiments (CERN 2000), CERN/LHCC, Oct. 2000, pp. 178-182, ISSN 0007-8328, ISBN 92-9083-172-3

[8] P. Leitao et al., "Test bench development for the radiation Hard GBTX ASIC," *JINST*, vol. 10, pp. 1-26, January 2015, doi: 10.1088/1748-0221/10/01/c01038.

[9] ALICE Collaboration, Radiation Dose and Fluence in ALICE after LS2, ALICE-PUBLIC-2018-012, 2018.

[10] C. Seguna, E. Gatt, G. Cataldo De, I. Grech and O. Casha "Proposal for a new ALICE CPV-HMPID front-end electronics topology," in 13th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME), June 2017, pp. 173-176, doi: 10.1109/PRIME.2017.7974135.

[11] C. Seguna et al., "A New FPGA-Based Controller Card for the Optimisation of the Front-End Readout Electronics of Charged-Particle Veto Detector at ALICE," The Second New Generation of Circuits and Systems Conference (NGCAS), November 2018, pp. 45-48, doi: 10.1109/NGCAS.2018.8572296.

[12] H. Witters, J. C. Santiard, Paolo Martinengno, "DILOGIC-2: A sparse data scan readout processor for the HMPID detector of ALICE," Proc. 6th Workshop on Electronics for LHC Experiments, Sep. 2000, pp. 179-183.

[13] F. Zhang et al., “Point-to-point readout for the ALICE EMCal detector,” *Nucl. Instrum. And Meth in Phys*, January 2014, pp. 157-162, doi: 10.1016/j.nima.2013.09.023.

[14] A. Velure, “Upgrades of ALICE TPC Front-End Electronics for Long Shutdown 1 and 2," IEEE Transactions on Nuclear Science, vol. 62, pp. 1040-1044, June 2015.

[15] ALICE Collaboration, Performance of the ALICE experiment at the CERN LHC, 2014.

Cartesian Systemic Emergence and its Resonance Thinking Facet: Why and How?

Marta Franova, Yves Kodratoff
LRI, UMR8623 du CNRS & INRIA Saclay
Bât. 660, Orsay, France
e-mail: mf@lri.fr, yvekod@gmail.com

Abstract— **Cartesian Systemic Emergence (CSE) is concerned with strategic aspects relative to the conception of Symbiotic Recursive Pulsative Systems intended to solve real-world problems handling control and prevention in incomplete domains. This work is performed to prepare fundamentals for designing automated tools that help to perform this complex task. This paper recalls the fundamental notions of CSE and presents the most important features of one particular way of thinking present in CSE. We call it 'Resonance Thinking'. Resonance Thinking takes care of generating and handling experiments during CSE. The work presented is related to systems design, cognitive, and computation models of human creative reasoning mechanisms as well as to the ML approach called "Ultra-Strong Learning" for computer-assisted learning of CSE and RT.**

***Keywords*— *Cartesian Systemic Emergence; Symbiotic Recursive Pulsative Systems; Resonance Thinking; systems design; implementation of human reasoning mechanisms; Ultra-Strong Learning*.**

I. INTRODUCTION

This paper presents one of the symbiotic parts, called Resonance Thinking (RT) [1], of our computer systems design theory of particular complex systems in incomplete domains. This theory of computer systems design is called Cartesian Systemic Emergence (CSE) and has been introduced in [16]. The complex systems concerned by CSE are called Symbiotic Recursive Pulsative System (SRPS, introduced in Section III.F).

The goal of CSE is to *formalize strategic aspects* of the human creation of SRPS. The originality of CSE consists in representing these strategic aspects as a *deductive-like problem-solving system* and on focusing on a formalization of *axioms* of such a deductive-like problem-solving system. Moreover, this formalization is performed in order to prepare fundamentals for *designing automated tools* that help humans, or even that are able alone to perform this complex task (as it would be convenient, for instance, for robots in space). RT is a process in CSE taking care of a generation of relevant experiments useful for pointing out the specifications for the necessary parts of the system to be constructed. Superficially, it might, therefore, be seen as a way that a system architect (see [7]) proceeds to an analysis of the system requirements in order to *decide*, which parts are necessary for constructing the corresponding system. This paper will show that CSE considers systems for which a simple analysis of system requirements is not sufficient and for which there must be a large specific experimentation phase leading to an on-purpose *invention* of the most parts of the system. Francis Bacon, in [2], calls 'experiments of Light' the experiments that have to lead to a specification of system parts and thus are related to the invention of the system axioms in contrast to the 'experiments of Fruit' that concern experiments that explore and exploit the axioms already given. RT can, therefore, be viewed as an example of a rigorous method for performing 'experiments of Light' in the context of CSE.

The purpose of this paper is five-fold:

- present the fundamental notions necessary for understanding RT and CSE;
- describe particularities of RT taking place in CSE;
- illustrate this method by a toy example, which nevertheless deals with a problem that many innovative researchers may have to face;
- suggest an application of Ultra-Strong Learning to a design of an evolving process for assisted teaching/learning CSE and RT;
- mention the main problems and challenges addressed by CSE and RT to various fields of Cognitive Science.

Since the context of CSE is rather unusual in multidisciplinary system design, in Section II, we describe this context. In Section III, we present the notions necessary for understanding the topic of CSE. Section IV presents a short description of CSE. Section V is concerned with a presentation of RT. Finally, Section VI brings forward some related work.

II. THE CONTEXT

In this paper, we deal with two paradigms in designing a problem-solving system. The first paradigm can be represented by the formula

$$\forall \text{ Problem } \exists \text{ System solves(System,Problem).} \quad \text{(P1)}$$

The second one can be represented by the formula

$$\exists \text{ System } \forall \text{ Problem solves(System,Problem).} \quad \text{(P2)}$$

(P1) states that for any problem Pb_i one can build at least one system (or a module) S_i able to solve Pb_i. (P1) leads to a library of particular heuristics or methods relevant to solving

individually each of Pb_i. (P1) is a *paradigm formula*, i.e., it has no truth value. It represents only a way to proceed when designing a system. Relying on (P1), one can, therefore, design a modular system S that is a modular composition of S_i that were previously built. Paradigm (P1) is useful when one of the main goals is to guarantee a simple maintenance of the resulting systems as well as a possibility of casual collaborations of the designers for each S_i [7] [28] [39]. Most system designing approaches are thus based on this paradigm. In this paper, the systems conceived via paradigm (P1) are called P1-systems. Similarly, the solutions to a problem conceived via (P1) are called P1-solutions. The same notation is used for P2-paradigm.

(P2) states that there exists at least one system that will solve all problems. The construction of a P2-system largely differs from that of a P1-system. The use of the P2-paradigm formula has to result in a single universal system S expressing the fact that this system represents a unique way in which all problems are solved. We can mention, for instance, the efforts of the approaches to Physics that tried to put in evidence one general theory of universe known as 'Theory of everything' [23]. The fact that, presently, there are two different results for macro and micro phenomena illustrates that a by cases analysis does not necessarily lead to a P2-system design. An illustration of a result of P2-paradigm is Peano's arithmetic for natural numbers (NAT). This example also illustrates that it is worthwhile to use P2-paradigm for the construction of systems that are not complete (in Gödel's sense [21]).

Since (P1) and (P2) are concerned with different goals, they are not competitive: each of them has its own particular 'competitive advantage' (see [38]) and these 'competitive advantages' are incomparable. Namely, as said above, P1-paradigm is useful when one of the main goals is to guarantee a *simple maintenance* of resulting systems. P2-paradigm is very useful for creating systems representing solutions for real-world problems that require handling *control* and *prevention* during the system design as well as in the resulting designed system. By control, we mean here the requirement to consider all the secondary effects of the evolution in the process of the construction of a particular system so that the constructed systems need *no future maintenance* (which guarantees, in fact, a non-obsolescence of the constructed systems), as it is, for instance, the case for Peano's axioms. By prevention, we mean here a careful anticipation of possible future practical needs, opening thus a way to a *smooth extension* of a previously, practically sufficient system. This can be illustrated by a smooth extension of NAT up to complex numbers. With respect to this particular competitive advantage of (P2), namely, handling control and prevention, there is a necessity to provide a formalization for what has been, so far, only an intuitive 'know-how' for designing P2-systems. This is the goal of CSE. CSE is a particular generalization of the experience that has been acquired while creating a reasonable solution for a real-world problem of implementing automatic construction of recursive programs specified by formal specifications in incomplete domains [14]. We shall refer to it as Program Synthesis (PS), for short.

A formal specification of a program (FSP) is a particular description of *what* the desired program P has to do. This description can, formally, be written as

$$\forall\, x\, \exists\, z\, \{\, IC(x) \Rightarrow IO(x,z)\, \}. \qquad (1)$$

Here x is an input vector, z is an output vector, IC is an input condition on x and IO is an input-output relation. A resulting program P is then a description of *how* it has to be done. In other words, in PS, the goal is to design a system S such that, for each formal specification FSP, the system finds a corresponding recursive program P that executes FSP. Another way of describing this is also to say that, PS aims at providing a reasonable solution for the formula

$$\exists\, S\, \forall\, FSP\ \text{'transforms'}(S,FSP), \qquad (2)$$

where 'transforms' means changing a 'what' (expressed by a formal specification FSP) into a recursive 'how'. It is thus clear that there is not only a close similarity between PS (expressed by the formula (2)) and P2-paradigm but also between FSP and P1-paradigm.

Our experience with the design of a PS-system required to express formally the thinking behind the development of our PS system. This motivated us to restrict our considerations of CSE to the design of systems that require a similar design process to that of our PS-system. In the next section, we present the fundamental notions that are necessary for understanding CSE as well as for a full specification of systems concerned (namely, SRPS introduced in Section III.F).

III. FUNDAMENTAL NOTIONS

The goal of CSE is to formalize strategic aspects of human creation of *informally* specified *symbiotic deductive-like problem-solving systems* in *incomplete domains* following our *pulsation* model. This formalization is performed in order to prepare fundamentals for *designing automated tools* that help humans, or even that are able alone to perform this complex task as already above stated. In this section, we recall five terms by which this goal is expressed and that will also be used in our presentation of Resonance Thinking, namely

- informal specification,
- incompleteness,
- symbiosis,
- deductive-like problem-solving systems, and
- pulsation.

The goal of CSE is to be considered in a P2-framework, i.e., CSE aims at a formalization that is a P2-system. Therefore, all these notions, that we need to define, are symbiotically interrelated. As a consequence, each of these fundamental notions cannot be clearly described without referring to the other fundamental notions. This is why, in

order to introduce such complex descriptions, we will present, at first, a rough description of their meaning independently of their aim to represent the basis of a P2-system. Such a rough description can also be used in the context of modular P1-systems.

The symbiosis of parts of a system means that, if even only one of these parts is eliminated, not only the system collapses but also all the other symbiotic parts collapse as well. An informal specification of a system is a description of this system that is somewhat vague, i.e., it may be unclear what the words in this description mean exactly. Deductive-like problem-solving systems are systems that are defined exactly by their corresponding axiomatic system. Incomplete domains are domains that are insufficiently formalized in the sense that there might exist several different interpretations corresponding to the considered formalization of the domain. Pulsation is a model for a particular kind of systems' evolutive improvement.

These notions were present in our work in their informal form from the start of our research on PS. In order to achieve an explicit formulation of CSE, we had to propose here a more formalized form of these notions.

A. *Informal specification*

The ***informal specification*** of the kind of systems that have to be constructed is a description of each system by a sentence in which occur terms that are underspecified, i.e., they are not yet exactly defined. When considered out of a particular context, the goal expressed by an informal specification may seem impossible to attain. For instance, let us consider two examples of informal specifications for real-world problems that may, outside a particular context, seem impossible to achieve.

(g1) In Computer Science, let us consider the goal 'Automate the construction of recursive programs via inductive theorem proving'.

(g2) In Cognitive Science, consider the goal 'Construct a scientific model of the human brain that solves all the questions and problems related to the brain mental processes'.

Both these goals seem impossible if we rely on the usual scientific contexts and the usual meanings of the terms in which these specifications are expressed. The notions and techniques introduced in this paper show that considering these goals from a different point of view opens a possibility to reconsider them as reasonable and achievable long-term scientific goals.

Even though from a management point of view, there may seem to be an unsolvable problem due to the informal specifications in which occur terms that are not yet exactly defined, from a scientific innovation perspective, it is acceptable that these terms evolve during the system construction. In other words, depending on some constraints and opportunities that may arise during the construction of a system, the meaning of the terms used in the starting specification will evolve and thus, the final delimitation of terms makes a part of the solution. In other words, the initial ambiguity of terms occurring in a given informal specification is eliminated by the provided solution. The evolution of these informal terms, as well as of the design of the system will then bring also an exact specification of the context to be considered.

The evolution process of NAT shows that NAT have been used with a rigor even before their final exact axiomatization by Peano (PAD). Therefore, in order to introduce a difference between *rigor* and *exactness*, in the framework of CSE, the notion of informal specification needs to be completed by the notions of formalized and formal specification. In CSE, ***formalized specification*** is an intermediary stage on the way from an informal to a formal specification. It consists of a collection of basic not yet exact definitions and basic not yet exactly defined tools that seem plausibly pointing out a successful completion process. Some inventive steps may still be needed to complete these inexact but rigorous tools so that their use and evolution through suitable experiences leads to their final form as well as to the final form of the basic definitions. In CSE, ***formal specification*** then consists of a complete solution represented by the working system (be it in its axiomatic form or in its implemented version), both a methodology for the system functioning and the complete knowledge necessary to the system construction. These all are needed in order to be used in further evolutive improvement - if such an improvement is relevant. Note that the notion of formal specification is here different from the notion of formal specification of a program (noted FSP in Section II). FSP describes what a program has to do, while in the context of CSE, a formal specification is a complete solution to a problem-solving task.

B. *Incompleteness and Practically Complete Systems*

As far as the notion of *incompleteness* is concerned, from a practical point of view, we know that full reality is unknown. What we may know at a given time can be formalized by an incomplete system.

In order to illustrate the informal (thus ambiguous) character of notions in incomplete theories, let us recall that, in a geometry obtained from Euclid's geometry by eliminating the postulate of parallels, a triangle can still be defined. However, in this incomplete 'theory', the sum of the triangle angles may differ from 180°. This means that the notion of triangle is incompletely (or inexactly or not clearly) defined in this particular purged (or mutilated) Euclid's geometry. In practice, it means that an informal definition covers several possible different interpretations of each 'defined' object. This illustrates what we mean by an informal definition. This also means that the completion process of a definition (its emergence) needs to orientate a choice – or rather, a construction – of an interpretation that is suitable for each particular problem to be solved. Such a choice, as well as the completion process, is guided by the formalization of objectives oriented towards a convenient solution of the informally specified problem. This means in

practice that, in any design or completion process, the goal is to formulate experiments oriented towards the construction of relevant constraints for the intended objective as well as for the final delimitation of notions.

We know that even though PAD defining NAT is an incomplete system [21], we all use successfully and with a rigor PAD in our everyday life as well as in exact sciences. This means that a practically incomplete system can already be used by all if we all learn to 'stick' to one exact interpretation. The situation is similar to the use of different geometries in real-world considerations where the experts of EG, of hyperbolic or of elliptic geometry stick rigorously to their completion of the above described 'mutilation' of EG. This means that it is usually the practical aims that point out towards the completion of an incomplete system.

As it is pointed out in [19], p. 20, it is meaningful to work with an incomplete system. An incomplete system that is useful to exploit despite its theoretical incompleteness is called, in our work, a ***practically complete system***.

From a decision point of view, it is well-known that incompleteness constitutes a large drawback. Incompleteness, however, is not at all a drawback for the practical purpose of solving real-world problems that are asking for some kind of invention [20]. This is because, from a construction point of view, incompleteness brings freedom for technological ingeniosity, resulting in possible new technological inventions. This means that changing the decision context, whenever possible or desirable, to the constructive action context (on-purpose designed for handling incompleteness) is a way to push us to think of a model for a process of a practically useful and theoretically reasonable completion of incomplete systems. Thus, this model should represent a kind of *directed anticipation* of future extensions leading to an ideal, maybe never achievable, complete system. In other words, it is interesting, for an incomplete system, to think of a practically useful rigorous process of completion and of evolutive improvement that would lead, at least by intention, to a complete system. Of course, such an intention would need to be specified, in advance, by an informal specification as is, for instance, expressed by the goal of CSE. In Section III.E, we present a model for such a process.

Let us note that, since an informal specification of a system contains terms that are not yet exactly defined, a particular informal specification points out to a context that can be represented by an incomplete environment. CSE can then be seen not only as a construction process for a system in its informally specified initial environment but also as a fruitful strategy for a progressive, evolutive, completion of this environment.

C. Symbiosis

When we handle incompleteness and informal specifications in the design of a system, we need to be aware of a particular interdependence of the parts of the resulting system. This is where symbiosis moves into the systems design methodology.

By ***symbiosis*** we understand a vitally separation-sensitive composition of several parts. By *vital separation-sensitivity* of a composition, we mean that eliminating one of its parts leads to three possible penalties. It may be a complete destruction or a non-recoverable mutilation or the uselessness of the remaining parts. This means that the widely used divide and conquer strategy, as well as analysis and synthesis, are not at all suitable when creating and extending symbiotic systems. Symbiosis is therefore different from the synergy that is a mutually profitable composition of elements that are not destroyed nor mutilated by separation.

The notion of symbiosis has been inspired to us by Descartes notion of 'conceptual distinction' [10], §62, p. 214 (in French: 'distinction par la pensée' [9], p. 131) and the notion of symbiosis used, a half-century ago, to describe the failed attempt to separate algae and fungi in lichen. This attempt was, at that time, a failure, since both separated parts died. We do not know whether it is today possible to separate these two parts. However, as far as we could consult the internet, modern science understands symbiosis only as a *balance* between symbiotic parts that can only be achieved by working together. In other words, it seems to us that modern understanding is far away from our perception of the parts possibly 'dying' by their separation.

Let us consider the following pictures.

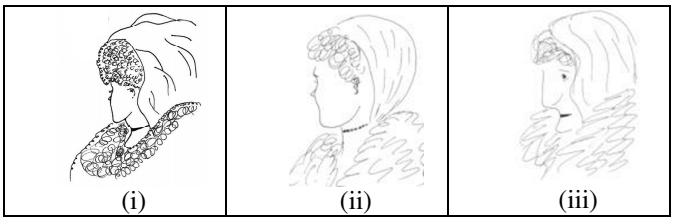


Figure 1. Example of pictorial symbiosis.

In Figure 1, a picture similar to (i) is known on the internet as 'young old woman illusion'. (i) can be seen as a symbiosis of (ii) and (iii). Indeed, if we remove from (i) the part that corresponds, for instance, to (ii), the part similar to the picture (iii), occurring previously in (i) before this step of removing, disappears as well. In other words, the result of removing one part in (i) is 'nothing'. This illustrates what we mean by 'destruction' in our definition of symbiosis.

Here, we need to point out that symbiotic parts do not necessarily need to overlap in the final symbiotic object. They may have a symbiotic, and maybe invisible, intersection that makes symbiotic their whole. From a systemic point of view, symbiosis of a system is embodied by the interdependence of all the notions and the parts of this system. For instance, a typical example of a symbiotic system is provided by PAD defining NAT. If we use the symbol ♦ for symbiotic compositions, a formal representation of the systemic definition of NAT reads:

$$NAT = 0 \blacklozenge Suc \blacklozenge NAT,$$

where Suc is the successor function. If we eliminate one part from this system, for instance 0, we may no more speak of Suc nor of NAT as such.

D. Deductive-like Problem-solving Systems

This section shows that there is a close relation of the notions of informal specification and of incompleteness to *formal* and *deductive-like problem-solving systems* (DPSS) specified initially by informal specifications. To any deductive-like problem-solving system corresponds its underlying deductive-like theory, a part of which constitute semantic definitions, non-abstract axioms, and 'intuitive' rules, as it is, for instance, in Euclid's geometry the complete description of which contains not only axioms but also the set of semantic definitions and real-world influenced practical rules (for instance, knowing *a priori* to draw a straight line between two points) necessary for handling the knowledge implicitly containing all the knowledge related to the use of this deductive theory. With respect to the existence of an informal specification for a DPSS, in our work, we point out the following difference between a deductive and a formal system.

A ***formal system*** is an *abstract* system where the axioms are given *without* the definitions containing the semantic of the objects considered and without the semantic rules that allow performing operations in this formal system. Moreover, when a formal system is considered in Science, its consistency is considered in terms of non-existence of a proof for a formula A as well as for the negation of A in this system. In contrast to this, a deductive-like system has a real-world model semantics.

As far as deductive systems are concerned, the history of the evolution of the formal axiomatic system for NAT shows that this formal system has been developed only at the end of the nineteenth century. In the previous centuries, even though semantically influenced intuitive 'definitions', intuitive 'axioms', and intuitive 'rules' were used and were evolving, these evolving intuitive objects were handled by researchers in a rigorous and intellectually proper way [19] [24]. This evolution process could then be concluded by the formulation of the final formal system used today, namely PAD. By ***deductive system*** we thus understand a system developed with a concrete real-world application as a model. This means that, in contrast to formal systems, the consistency of deductive systems in their evolutive process of construction is proved by the existence of a concrete model. Therefore, a deductive system is in our work viewed as a result of the development of a relevant, possibly incomplete system of semantic axioms for a particular intended application interrelated with semantic definitions of objects that occur in these semantic axioms and semantic rules. In the final stage of the development, when exclusively the manipulation (and not construction or completion) purposes can be and are considered, a deductive system can be viewed as a formal system obtained by abstraction, however, its completeness is not viewed from a theoretical point of view but from the point of view of practical completeness.

E. Pulsation

Pulsation is a model for construction and evolutive improvement of incomplete, but practically complete systems that are concerned with the above-described factors of control and prevention (see Section II). In other words, pulsation provides a rigorous framework for the completion process of incomplete systems. This model relies on our particular handling of Ackermann's function. We shall recall now the features of its handling that will also be referred to later in the paper.

Let 'ack' be Ackermann's function defined, as in [40], by its standard definition, i.e.,

$$ack(0,n) = n+1 \quad (3)$$

$$ack(m+1,0) = ack(m,1) \quad (4)$$

$$ack(m+1,n+1) = ack(m,ack(m+1,n)). \quad (5)$$

Since ack is a non-primitive recursive function, by definition of non-primitive recursion, it is a particular composition of an infinite sequence of primitive recursive functions. In [18], it is shown how, for computing the value of ack(a,b), for given a and b, one can replace the standard definition of Ackermann's function by an on-purpose recursive macro, which consists of a finite sequence f_n of primitive recursive functions the "infinite limit" of which corresponds exactly to ack, i.e.,

$$\lim_{n\to\infty} f_n = \text{ack}. \quad (6)$$

This trick changes the non-primitive recursive computation of ack(a,b) for particular given values of a and b to primitive recursion and thus makes Ackermann's function suitable as a model for practical purposes of constructing systems that are not dealing with the computations on NAT but with solving real-world problems. In similarity to the infinite sequence, which is used in [18] to construct ack, the evolutive improvement (i.e., pulsation), relies on a construction of a potentially infinite sequence of systems that might, in an ideal world, be used to construct a global *'Ackermann's system'* that contains all of these systems. In our work, by pulsation we thus understand a progressive construction of a potentially infinite sequence T_0, T_1, …, T_n, T_{n+1}, … such that

- T_0 is the initial informal specification,
- T_i, for $i > 0$, is an incomplete, but a practically complete deductive-like system,
- $T_i \subset T_{i+1}, T_i \neq T_{i+1}$ (for i = 0, 1, 2, …), and
- an infinite limit of this sequence represents an ideal, complete deductive system S.

We say here that T_{i+1} is a *practical completion* of T_i (for i = 0, 1, 2, …). We have

$$T_{i+1} = T_i \cup A_i, \quad (7)$$

where A_i is the set of axioms corresponding to the extension of T_{i+1} from T_i.

The fourth requirement can be written as

$$\lim_{n\to\infty} T_n = S, \quad (8)$$

where S is the ideal, complete Ackermann's system corresponding to the initially given informal specification.

The notion of Pulsation was inspired by the informal notion of 'gradual building of theories' presented in [19], p. 21, and the informal notion of 'pulsative system' presented in [12], p. 165. With respect to the constraints of handling control and prevention in the design of a DPSS as well as in a concrete DPSS itself, we have just developed and improved the formalization of these notions into our notion of Pulsation using Ackermann's function as a model. We have done so since this function models 'thinking of everything' (it is thus related to P2-paradigm and similar to the above mentioned "Theory of Everything" in Physics) as well as an effort that leads to achieving this goal.

Pulsation does not reduce to one particular step in the sequence T_0, T_1, …, T_n, T_{n+1}, …. This means that pulsative systems are formalized progressively and potentially indefinitely. Thus, this new paradigm of the pulsative evolution of systems corresponds well to Bacon's understanding of Progress as a result of the work of several generations [27].

Pulsation has been introduced in [18]. Pulsation is a model that does not describe how the particular systems in the sequence T_0, T_1, …, T_n, T_{n+1}, … are constructed. This is the role of Cartesian Systemic Emergence described below in Section IV.

F. Putting it together

Let us consider an example of a symbiotic system for which we put symbiosis together with Pulsation and P2-paradigm. Thus, let us consider the above informal specification (g2) from Cognitive Science. Now, we may ask whether a solution might not be something like

$$\lim_{n\to\infty} Brain_n = Brain, \quad (9)$$

where

$$Brain_n = Left\_Brain_n \blacklozenge Right\_Brain_n \blacklozenge RNK_n \;\&$$
$$Left\_Brain_n = Brain_{n-1} \blacklozenge RNK1_n \;\&$$
$$Right\_Brain_n = Brain_{n-1} \blacklozenge RNK2_n$$

Here, RNK_n is a new knowledge related to symbiotic composition of left and right brains in the n-th pulsation step, $RNK1_n$ and $RNK2_n$ are relevant new knowledge extending, by the process of practical completion, the previous knowledge respectively about $Left\_Brain_{n-1}$ and $Right\_Brain_{n-1}$. This means that, instead of studying the brain as a synergy of two elements (left and right brain) [3], it might be interesting to explore the potential of this new symbiotic paradigm. Note also that (9) represents a recursive model of the brain, which explicitly indicates its evolutive character. This example also illustrates that, from a systemic point of view, symbiosis of a recursive system is embodied by the interdependence (explicit or implicit) of all notions and parts of this system.

Above, we have introduced the fundamental notions that contribute to understanding CSE and ST. These notions allow us to introduce, in multidisciplinary complex systems design, the notion of *Symbiotic Recursive Pulsative Systems* (SRPS) that are, by definition, systems that are implicitly or explicitly symbiotic, which are recursive either by systemic recursion or by the process of evolutive improvement via Pulsation, and that are pulsative, whenever the model of Pulsation (together with the notion of practical completeness) is used in their design. In order to point out that SRPS are developed via paradigm (P2), we may call them P2-SRPS.

We thus come to Cartesian Systemic Emergence and its facet 'Resonance Thinking' that takes care of generating and exploitation of some experiences in this process of CSE.

IV. CARTESIAN SYSTEMIC EMERGENCE

The above-introduced vocabulary allows us now to say that the goal of CSE is to *formalize strategic aspects* of human creation of P2-SRPS in order to allow a collaboration in P2-SRPS-projects and to develop automated tools that help human in this task or even that perform this task of P2-SRPS-creation themselves. As presented in [16], the main features of CSE are as follows:

- It works with an informally specified goal.
- It handles incompleteness.
- Takes into account symbiosis and pulsation.
- Generates experiences.
- 'Oscillates' between the paradigms (P1) and (P2) in order to reach a solution described by (P2).

There are four fundamental symbiotic facets of CSE:

(a) Pulsative Thinking, i.e., taking care of security, control, and prevention.
(b) Metamorphic Thinking, i.e., taking care of resulting epistemological equivalence between P2-paradigm and particular CSE-handling P1-paradigm.
(c) Symbiotic Thinking, i.e., taking care of the construction of a symbiotic system.
(d) Resonance Thinking, i.e., taking care of generating and handling experiments.

In [1], we show that these four rules are inspired by Descartes' method [8] considered in symbiotic environments.

As we can realize while trying to give an *exact* description of the old-young lady picture given in Figure 1.i, a description of one part in a symbiotic composition (such as 'old lady' in Figure 1.iii) is not a simple task. Indeed, while considering Figure 1.i, an exact description of the old lady (as in Figure 1.iii) would imperatively require explicit references to the young lady (as in Figure 1.ii). Therefore, in this paper, we do not intend yet to provide a complete description of Resonance Thinking, since we need first to describe more in details Metamorphic Thinking (MT) and Symbiotic Thinking (ST).

The next section presents the main ideas necessary for a rough understanding of RT.

V. RESONANCE THINKING

Experiences represent an important part of each real-world complex system designing process. For P1-paradigm, with respect to the possibility of a heuristics library representing a set of partial solutions for different types of problems, these experiences are mostly clever combinations of the already acquired knowledge. However, by the requirement of a universal, 'unique', character of the solution for P2-paradigm, it is necessary to perform experiences that are relevant to the invention process of the knowledge that is not yet available and that is not a combination of the available knowledge. In other words, since CSE aims at a construction of a deductive-like problem-solving P2-system, the ultimate goal of RT is to generate experiences that suggest the essential symbiotic parts (understood as primitive notions) of the desired P2-system. As a hint, we could here recall that Francis Bacon, called 'experiments of Light' such experiences.

In Section II, we explained what is the difference between the use of (P2) and (P1). RT exploits the idea of 'oscillating' between (P2) and (P1) in order to come to a global P2-solution by studying particularly chosen P1-experiments. In order to be fruitful and justified, a switch from (P2) to (P1), i.e., an 'oscillation' step, generally has to rely on the above mentioned symbiotic part of CSE, namely MT. Roughly speaking, MT takes a care of a rigorous, epistemologically and pragmatically justified transformation of paradigm (P2) into the context of paradigm (P1). Such an epistemic justification relies on particular features of 'case analysis' used while developing a P2-system. As we have said above in Section II, not all kinds of 'case analysis' allow coming from P1-systems to a universal P2-system. Therefore, in our future work, we aim to explain in detail what makes our 'case analysis' reach the desired goal. However, we need to present ST first, since all the symbiotic parts of CSE have to be included in such an explanation.

In other words, MT provides a switch from (P2) to (P1) that is useful in order to generate experiments leading, within the P1-framework, to some hints and inspiration for solving finding a P2-solution. This means that the goal of RT in the construction of a P2-system is twofold:

- to hint at the parts of the future P2-system,
- to provide a guarantee that the suggested parts are symbiotically interrelated.

The hints and inspirations resulting from an achievement of the first goal represent temporary underspecified constraints that enlarge the already existing set of underspecified constraints.

We shall present RT and its basic notions with the help of a toy example used, in [16], for illustration of CSE. In comparison to examples provided by PS-framework, this example is simpler and could illustrate many other scientific fields than PS-research does. The example problem presented here concerns conveying a new original scientific knowledge in such a way that its essential content and creative potential are preserved by the next generations. This is not a trivial problem as already pointed out in the past [2] [8]. Our experience confirms that, for new knowledge that concerns the creation and a smooth extension of symbiotic recursive systems, this problem remains relevant until now.

A. *Specification of a toy example*

In this section we present the context of our example illustrating RT.

Let us suppose that René is a founder of a novel scientific theory with a high pulsative potential. Referring back to the founders' unhappy past experience, he (our 'René') needs to ask himself: "How to build some 'works' able to convey the full complexity of my new theory while simultaneously preventing a degradation of its pulsative potential?" In a more formal way, René must solve a problem informally specified as:

$$\begin{aligned} &\exists \text{works } \forall \text{disciple conveys(René, works) \&} \\ &\quad \text{conveys(works,disciple)} \Rightarrow \\ &\text{essential\_of(René) = essential\_of(disciple).} \end{aligned} \tag{10}$$

Note that this problem has the same logical structure as P2-paradigm. Specification (10) is an informal specification. As said above, this means that the notions that appear in (10) are not defined in a rigorous way. They are only specified in an informal way in terms of some non-formal criteria (i.e., a kind of underspecified constraints). This means that a solution 'works' for (10) has to emerge simultaneously with suitable formalizations (thus, the final definitions) of notions that occur in (10). In the following, we shall denote by D_t the set of (initially underspecified) sentences specifying 'to convey' and by D_e the set of (initially underspecified) sentences specifying 'essential_of'. These sets evolve in the process of CSE and Resonance Thinking towards a more rigorous final form. For the presentation simplicity, we do not involve such an evolution in our notation.

B. *Resonance Thinking in Action*

In order to solve (10), we perform a particular switch to a framework of experiments described by the formula

$$\forall \text{disciple } \exists \text{works conveys(René, works) \& conveys(works,disciple)} \Rightarrow \text{essential\_of(René)} = \text{essential\_of(disciple)}. \quad (11)$$

This formula represents P1-paradigm. Above, we have explained that there is a difference between the use of (P2) and (P1). This obviously applies to their instances (10) and (11). Above we have explained also that the goal of RT in a construction of P2-system is twofold, namely to hint at the parts of the future P2-system and to provide a guarantee that the suggested parts are symbiotically interrelated.

As far as the first goal is concerned, in order to generate such inspiring experiments hinting the parts of the future system, while considering (11), from the set of all disciples, we chose a finite number of disciples d_0, d_1, …, d_n that seem highly different from each other so that each of them seems *a priori* to need a different 'works'. We shall call *representatives* these disciples. In other words, our experience shows us that various challenging experiments are needed to obtain some inspiration contributing to a solution of (10) in the framework of the P2-paradigm. Note that we are working here in the context of real-world situations (the semantics of which is well-known) where we consider the representatives for which it is meaningful to suppose that some solutions can or should be found in the framework of the P1-paradigm. This is related to the notion of practical completeness of the resulting system. In other words, representatives describe situations that are difficult but not *a priori* unsolvable ones. Note also that we order these disciples in a numbered sequence just for the presentation purposes. This will be useful when describing recursive procedures that handle this finite set of disciples.

Recall that the two operators 'conveys' and 'essential_of' are here informally specified only by some set of sentences that represent informal descriptions (i.e., underspecified constraints) relative to these notions. Thus, we shall replace these notions by their informal descriptions. Above, we have denoted by D_t the set of sentences specifying 'to convey' and by D_e the set of sentences specifying 'essence_of'. Therefore, (11) writes as

$$\forall \text{disciple } \exists \text{works } \{ D_t(\text{René, works}) \ \& \ D_t(\text{works,disciple}) \Rightarrow D_e(\text{René}) = D_e(\text{disciple}) \}. \quad (12)$$

Let us consider (12) for each particular d_i, i.e.,

$$\exists \text{works } \{ D_t(\text{René, works}) \ \& \ D_t(\text{works}, d_i) \Rightarrow D_e(\text{René}) = D_e(d_i) \}. \quad (13)$$

For a moment, let us suppose that a solution for (13) is found for each d_i while, during this 'search', oscillating between paradigms (P1) and (P2). In this case, the oscillation means that while, in (13), we are working in the context of (P1), we seek for solutions that are not the results of clever heuristics but are the results of trying to capture 'parts' of a general method that might be a basis for a P2-system. This means that we 'switch' mentally from (P1) back to (P2). In other words, while seeking a P1-solution we, in fact, aim at a P2-solution. Practically, this manifest by the fact that one is aware of the danger that comes from the attraction of clever solutions. Clever P1-solutions are usually a barrier to a discovery or invention of a unified way of solving problems, which is aimed at by (P2). The actually obtained (almost P2-like but in reality a) P1-solution consists of a concrete value w_i for 'works' and of less informal descriptors $D_{t,i}$ and $D_{e,i}$. We shall note $Sol_i = \{ w_i, D_{t,i}, D_{e,i} \}$. Due to a careful oscillation between paradigms (P1) and (P2), the descriptors $D_{t,i}$, $D_{e,i}$ and w_i refine 'works' and the operators 'to convey' and 'essential_of' in (10). These resulting refinements have to 'resonate' with the framework of paradigm (P2). By their resonating we mean that, during the experimentation process, we need to feel that they might, probably after some 'judicious adaptations', be applied also to other instances of 'disciple'. In other words, the parts suggested by d_i have to be symbiotically compatible with the parts suggested by d_j (for i, j $i \neq j$). Therefore, while the first step of RT (taking into the account the first goal of RT) lies in a careful choice of representatives leading thus to specific experiments, the second step of RT takes care of the second goal, namely it generates experiences that have to provide a guarantee that the suggested parts of the future system are symbiotically interrelated. These steps are interrelated in the sense that when a symbiotic interrelationship is incompatible among some parts suggested by d_i and d_j (for i, j $i \neq j$), new representatives are chosen and the consideration of the failure representative d_f (i.e., a representative suggesting incompatible parts) is postponed until more experience allows to suggest another solution Sol_f for d_f.

Let us proceed now to the second goal of RT, i.e., providing a guarantee that the suggested parts are symbiotically interrelated. It is important to note here that RT relies heavily on 'resonance' as defined by: "a quality that makes something personally meaningful or important to someone" (e.g., as in Merriam-Webster Dictionary). The second step of RT thus involves the ability to create and explore personally meaningful or important relations in the process of generating and handling experiments.

We have seen that the first goal is approached via a choice of relevant representatives, i.e., a *choice* of relevant experiments. The second goal is approached via a particular process of *generating* and *handling* complementary experiments. We are going to describe it in the framework of René's example.

At this stage, we suppose that (13) for d_0 is already solved. Sol_0 represents a 'temporary' solution for d_0. By 'temporary' we mean that this solution will still have to be approved or modified by RT. Procedurally, the part of generating experiences of RT is based on two procedures for which we cannot yet provide a detailed description (thus, making explicit also 'handling experiments' part of RT), as they rely also on other symbiotic facets of CSE not introduced yet (namely, MT and ST mentioned above). We shall, therefore, concentrate on explaining the role of these procedures. The first procedure will be called *topological*

symbiosis (noted *ts*) and it is also a primitive operation for the second procedure. The second procedure is called *complementary topological symbiosis* (noted *cts*). Both of these procedures require creativity in developing symbiotic systems. They are therefore to be handled, for the time being, by a creative human person. The following description of the role of *ts* and *cts* illustrate some of the challenges that *ts* and *cts* have to tackle.

1) On symbiosis in Resonance Thinking

We need to point out here two particular features of *ts*. The first one concerns the character of possible "mutilations" performed by *ts* and the second one concerns its goal.

In Section III.C, we have presented an example of a pictorial symbiosis of two different women. One woman is young, the other is old. The resulting symbiosis is a face that can be seen simultaneously as a young and an old woman. The original two pictures of women have to be 'mutilated' so that the resulting symbiotic picture is a convincing illusion. For instance, an eye of the old woman and an ear of the young woman overlap in the symbiotic picture. As for the opposite ages of the women on the initial pictures, they are 'merged', since the symbiotic picture is at the same 'old' and 'young'.

A systemic symbiosis manifests itself not so much as 'merging' contradictory facets of the considered system (as 'merging' two opposites, namely young and old in the above mentioned pictorial symbiosis), but as constructing an emergent vitally separation-sensitive interdependence (i.e., symbiosis) of suggested parts of the system. Thus, while the first goal of *ts* refers to a seemingly useless 'mutilation' of suggested parts of the system, the second goal of *ts* expresses the importance of such a mutilation in a search of relevant vitally separation-sensitive interdependence (i.e., symbiosis) of suggested parts of the system. Note that these goals are feasible thanks to the fact that the suggested parts are informally specified, thus 'temporary' and evolving, till the end of the process of the system construction. By 'temporary' we mean that these parts will still have to be approved or modified by CSE and RT.

2) On generating experiments in Resonance Thinking

We are going to describe *ts* and *cts* in the framework of René's example. At this stage, we suppose that (13) for d_0 is already solved. This provides the solution Sol_0 for d_0. Sol_0 represents a 'temporary' solution for d_0. Similarly, for other disciples d_1, …, d_n, we will obtain Sol_1, Sol_2, and so on. We assume here that the solutions are obtained in a particular 'linear' way, one after another. This 'linear' way looks as follows.

Once Sol_0 is constructed, a 'temporary' solution Sol_1 for d_1 is constructed ('temporary' in the same way as Sol_0 is a 'temporary' solution for d_0, i.e., they will have to be still approved or modified by CSE and RT). Note that both these constructions may lead to new experiences and thus, *they may modify the initial environment* by refining the informal notions of our definition (10) of our problem. For the sake of simplicity, we do not describe explicitly below this evolution of the environment, though we take it into account by calling it a 'feedback' when we use it.

Now, let us suppose that we have solved the problem for d_1. Before starting solving the problem for the next one, we try to take into account the informal notions present in (10). This try amounts to an attempt to 'merge' the solutions Sol_0 and Sol_1 using topological symbiosis *ts*, i.e., we try to achieve their symbiotic composition that resonates (as explained above) with the informal specification (10). We shall denote this process by $ts(Sol_0,Sol_1)$.

If solving $ts(S_0,S_1)$ fails, i.e., we cannot find relevant refinements, we keep in mind the feedback obtained while constructing Sol_0 and Sol_1, as well as the failure reasons of $ts(S_0,S_1)$. This failed step will have to be redone later while relying on some inspirations that may arise while finding the solutions for the next disciples. If this process fails, the problem will have to be considered as a challenge for one of the next pulsation steps.

If the process $ts(Sol_0,Sol_1)$ succeeds, both solutions are temporarily approved. Then, keeping in mind all the feedback obtained, a solution of (13) for d_2 is constructed. One might suppose that this process may continue linearly as suggested by its beginning, as we just have seen. However, recall that we work in an environment that requires control and prevention. This means that generating complementary experiments for the topological symbiosis of solutions constructed is necessary. We call *complementary topological symbiosis* (noted *cts*) this procedure for generating new experiments.

Roughly speaking, *cts* is a particular generation process (defined with the help of *ts*) for creating experiments. The goal of these complementary experiments is to provide inspirations for further refinement of underspecified notions and constraints. Similarly to the computation of ack (see [17]), in the process of generating experiments (via *ts*) for Sol_m and Sol_n, i.e., while 'computing' $cts(Sol_m,Sol_n)$, the operation $ts(Sol_i,Sol_j)$ for other solutions Sol_i and Sol_j has to be performed several times.

Let us denote by $ts_1(Sol_i,Sol_j)$ the solution of the first computation, by $ts_2(Sol_i,Sol_j)$ for the second computation, and so on. It is important to point out that $ts_p(Sol_i,Sol_j)$ and $ts_q(Sol_i,Sol_j)$ in this sequence of computations may carry two different feedbacks. Indeed, each inner step of *cts* (i.e., evaluating $cts(Sol_m,Sol_n)$), may bring new refinements, constraints as well as it may point out to missing knowledge or second-order notions and procedures. The procedures *ts* and *cts* have to ensure that not only reasonable and achievable solutions are obtained but that a possibility of future evolutions are guaranteed while properly handling prevention and control.

We do not present here an algorithm for *cts* since an automated execution of *cts* is not yet solved. However, based on our large experience, already now we may express that an important requirement of this procedure is that, for a computation of $cts(Sol_m,Sol_n)$, it must contain considering $ts(Sol_m,Sol_n)$ in randomly generated environments provided by the computation of $cts(Sol_p,Sol_q)$, where ($p < m$ and $q < n$) or ($p = m$ and $q < n$) or ($p < n$ and $q = m$) and $ts(Sol_p,Sol_q)$

has to be performed several times for some p and q. In other words, *cts* is an environments generation process that has to be inspired by the computation trace of Ackermann's function [17]. Recall that, in RT, *ts* and *cts* are, in our case, presently performed by a human mind. This means that human mind can rely on relevant creativity in order to decrease the number of repetitions. In consequence, even though *ts* and *cts* are not simple, CSE and RT are not overwhelming tasks for human performers. However, they may be overwhelming for a human observer even in this simplified form.

VI. DISCUSSION/RELATED WORK

Usually, in the design of systems, there exists a clear distinction in the roles of a system architect/analyst, of a system designer and of system integrator (see [7]). This is not the case in the design of SRPS systems. Namely, these particular systems design requires 'one mind' for performing these tasks simultaneously (i.e., symbiotically). Moreover, collaboration on the design of SRPS systems is a scientific work that can itself be characterized as a complex emergent system [41]. In Section VI.A, we illustrate some negative consequences of an attempt to simplify and replace a symbiotic process by a process that should simulate a synergic collaboration. In Section VI.B, we describe our study of an interesting relationship between mental processes of Resonance Thinking and the so-called Ultra Strong Learning. Section VI.C brings some insights on a possible influence of CSE on some research topics in the field of Cognitive Science. Sections VI.D and VI.E present a comparison of our research, General System Theory and Multidisciplinary Systems Design, respectively.

A. On Simplification and Delegation

As said above, CSE and RT are concerned with the human creation of symbiotic systems. One may, therefore, wonder whether this process of creation cannot be described in a simple way, so that the usual process of delegation and synergic collaboration might be used. That such an initiative is not at all reasonable in the case of symbiotic objects may easily be illustrated by a study of the creation of Figure 1.i. Namely, it is not difficult to foresee problems if we charge several skilled painters that are not familiar with this picture to come out with a solution for the problem: Create a picture that makes some people to see exclusively a young woman, some people exclusively an old woman and some people both the women. Of course, Figure 1.i. is a convenient solution in this case. While this illustration is short, it does not rely on a scientific study. Such scientific study should justify the hypothesis that

- the process of creation of practically complete symbiotic systems cannot be simplified, and
- the usual synergic delegation cannot be used.

In order to give an illustration that is, in our opinion, scientifically admissible, let us consider, side by side,

- the above mentioned standard definition of Ackermann's function; we have used the name ack for it, and
- an unusual definition of Ackermann's function; we shall use the name ak for it.

Thus, we have:

$$ack(0,n) = n+1 \quad (14)$$

$$ack(m+1,0) = ack(m,1) \quad (15)$$

$$ack(m+1,n+1) = ack(m,ack(m+1,n)) \quad (16)$$

and

$$ak(x,0) = sf(x) \quad (17)$$

$$ak(x,y+1) = ak(x,y) + sf3(x,y), \quad (18)$$

where sf is defined by

$$sf(0) = 1 \quad (19)$$

$$sf(a1+1) = sf(a1) + sf1(a1). \quad (20)$$

Here, sf1 is defined by

$$sf1(0) = 1 \quad (21)$$

$$sf1(b+1) = sf2(b,sf(b) + sf1(b)) \quad (22)$$

and, sf2 is defined by

$$sf2(0,y) = 1 \quad (23)$$

$$sf2(a+1,0) = 1+ sf2(a,1) \quad (24)$$

$$sf2(a+1,b+1) = sf2(a+1,b) + sf2(a,b+sf2(a+1,b)) - 1 \quad (25)$$

Finally, sf3 is defined by

$$sf3(0,y) = 1 \quad (26)$$

$$sf3(a+1,y) = sf2(a,ak(a+1,y)) \quad (27)$$

We have shown, in [13], in a constructive way, that ak is computationally equivalent to ack, i.e., it implements the program ack, even though it does so in a different form. Thus, both ack and ak are non-primitive recursive programs. While ack is defined recursively with respect to the first argument (and by-cases with respect to the second argument), ak is defined recursively exclusively with respect to the second argument. We shall call ack-form the definition of ack in terms of (14), (15) and (16). We shall call ak-form the definition of ak in terms of (17) and (18). Similarly, sf-form stands for the definition of sf by (19) and (20), sf1-form stands for (21) and (22), sf2-form stands for (23), (24) and (25), and finally, sf3-form stands for (26) and (27).

Pragmatically speaking, we have said, in [18], that Ackermann's function can be seen as 'thinking of

everything' for a given problem. While some may argue that this is impossible because of the incompleteness of reality, our model of pulsation (see Section III.E) illustrates that this can reasonably be done if we accept to work in the framework of a potentially infinite sequence of practically complete systems.

Let us consider ack-form and ak-form from the management point of view of simplification and delegation. If we consider a name for a program as a person charged to perform the task of this program, the first and second argument as the right hand and left hand, respectively, we may notice that

- ack uses both hands and executes the task alone;
- ak uses just left hand, delegates completely to sf the computation of (17) and, as for the computation (18), ak together with sf3 work on it.

From the point of view of standard management, we may note that ak uses only the left hand – being thus able to accept other tasks of company employing him – and that he also delegates a full case (17) to sf. Therefore, his move – namely, to delegate fully one task and to collaborate with sf3 on the second case – are highly appreciated. Using the standard criteria of management ak is considered as a better manager. We may say even that ak seems much clever than ack since he has only two lines to describe his work, whereas for ack we count three lines. Moreover, ak uses a description of his work (i.e., ak-form) that looks very 'nice' from the perspective of the form of primitive recursive functions. Let us recall that primitive recursive functions are a composition of a finite set of functions.

Before continuing further, let us recall that it is not unusual that we feel at ease with forms that look familiar at first sight (since they look 'nice') and then we do not look too much at details [26]. We shall not skip the details in this case, and we shall go deep inside.

First, let us denote by BC_{ak} the set of members involved while solving the base step of ak-form, i.e., in (17). We have BC_{ak} = {sf, sf1, sf2}. In this set, sf is, from the management point of view, on the same level as sf1, since sf, in (20), collaborates with sf1 and sf1, in (22), collaborates with sf. Then, we have that sf1 is on a higher level that sf2, since sf1 partly delegates, in (22), to sf2. We can note that sf-form and sf1-form look 'nice' from the perspective of the form of primitive recursive functions. sf2 works alone with the work that is charged to him by his superiors.

Let us denote GC_{ak} the set of members involved while solving the general step of ak-form, i.e., in (18). We have GC_{ak} = {ak, sf3, sf2}. We can also note that ak collaborates with sf3, see (18). Finally, sf3's work is not recursive, in (27), he calls ak for help and then charges sf2 to do the final work. Since sf2 occurs in both sets BC_{ak} and GC_{ak}, since it is given a work by his hierarchic superiors, i.e., it is a simple auxiliary, it might logically seem that his work is very insignificant. However, if we compare sf2-form and the original ack-form (i.e., (14), (15) and (16)), we can see that sf2-form is as 'difficult' to appreciate - from the simplicity and 'niceness' point of view – as ack-form. In the language of mathematics, sf2 seems a non-primitive recursive program as it is the case for ack. Since there is a computational equivalence of ack and ak [13], sf2 must be the program that guarantees this non-primitive recursive character of ak. While we have shown [15] that the non-primitive recursion character of ack can be proved in a simple constructive way, in order to prove the non-primitive recursive character of sf2, one needs to return to [40] to seek inspiration for a proof by contradiction. Indeed, a proof by contradiction is usually presented to prove the non-recursive character of ack. And then, since ak uses sf2 in its computation, ak is obviously a non-primitive recursive program, since a primitive recursive program is, by definition of primitive recursion, a composition of a finite set of primitive recursive programs. Finally, while ack is computed in the base step, i.e., in (14), by a primitive recursive program n+1, ak computes its base step relying on non-primitive recursion of sf2.

In other words, from the point of view of simplification and delegation the program ak, with all his necessary auxiliaries sf, sf1, sf2 and sf3, gives an illusion of simplification and delegation while decreasing the computational efficiency of ack. In other words, even though ack works alone, he is more efficient than ak with his collaborators and auxiliaries.

Our illustration in this section shows that a synergic collaboration is unsuitable for the projects of the creation of SRPS. These projects require 'symbiotic management' and 'symbiotic collaborations'. These topics will be addressed in our future work.

B. Topological Symbiosis and IML

In the process of RT, performing topological symbiosis gives a rise to several problems that are also met in the field of Inductive Machine Learning (IML), as illustrated below.

In this research field, it is interesting to make the difference between several 'levels' of learning and Michie [32] provided three criteria for the evaluation of a degree of 'value' for Machine Learning (ML) results. He classified them as weak, strong and ultra-strong criteria. For him, the weak criterion identifies the classical case where the machine learner produces improved predictive performance with increasing amounts of data. The importance of this 'weak' ML is illustrated by the large success of what is nowadays called "Artificial Intelligence."

The two other criteria may nevertheless be the root for even more powerful techniques of learning.

'Strong' ML generates also some kind of symbolic explanations enabling the human persons receiving the results of a strong ML system to understand the why of these results provided the machine. This differentiates two subfields of ML, namely one based on numerical computations and one based on symbolic lines of reasoning.

Finally, Michie's Ultra-Strong Learning (U-SL) implies the existence of a kind of collaboration (not necessarily a symbiotic one) taking place in between a machine and a human in a way such that, at first, the ML system teaches some valid information to his human user, as may do IML applied to a specific training data set. In order to reach the

ultra-strong level, human performance (on the same training data) has to be proven becoming more efficient than the one obtained by human studying the training data alone.

This last requirement asks for a computer system to perform three complementary abilities, as is shown in [33] [34].

The first one is to *generate pieces of programs* that are 'immediately' understandable to a human being. Since the manipulated programming language is Prolog, Muggleton's experiments have been carried on people who underwent at least two terms of Prolog teaching.

The second one (called *'Step A'* by Muggleton) is that the program, while it is running in order to answer a question, is able to generate new Prolog clauses. This ability is called "*predicate invention*" since each Prolog program is built with a concatenation of such predicates. The topic of predicate invention has been a basic problem for many Prolog specialists since the beginning of this research field. Fulfilling this condition amounts to achieve Michie's condition for a "strong machine learning" system. Muggleton's approach to this problem can be summarized as follows [33]: the system makes use of a controlled pattern matching of a higher order knowledge, provided in the form of meta-knowledge handled by a meta-interpreter. New knowledge is obtained by proving that a meta-goal is valid on a selected set of true examples. Note that this procedure is not submitted to our constraints of symbiosis and pulsation.

In our presentation, Muggleton's Step A. can be seen as a partial instance of what we call here *ts*, the role of which is to create new relationships induced from the data.

The third needed ability is stated below, in Muggleton's step B.

Muggleton's step B. Once new rules are found during Step A., [34] makes use of these rules in order to *select a set of significant examples*. This selection could work in a random way, generating a random mixture of examples illustrating both the old rules and the new generated one. In Muggleton's context, these examples are actually generated in such an order as to constitute the 'background knowledge' provided to a human learner. Thus, some selection among the possibly generated rules has to be done in order to be sure to obtain a 'significant' background knowledge.

In our presentation, Muggleton's step B. can be seen as a partial instance of what we call here *cts*. The goal of *cts*, similarly to step B., is to select a set of examples in order to complete or to enlarge the new knowledge initially generated by *ts*. However, in difference with step B., *cts* generates random examples because this provides a greater probability of generation of new (useful or missing) knowledge. This is coherent with our choice of a set of disciples for which solving (13) (see Section V) is rather difficult. We have mentioned above that this necessarily leads to a need for greater creativity and thus leads more efficiently to practical completeness of resulting system, here 'works' as in formula (10). At this stage, we already can acknowledge that Muggleton's steps A. and B. might be used as an inspirational model for programming the main procedural features of *ts* and *cts*.

As far as further steps in Muggleton's approach are concerned, his purpose is to get a set of rules such that their knowledge will improve a human's programming behavior. His success in this task, as explained in [34], proves that Muggleton's work is a success in implementing the very first ML program fulfilling the 'extra-strong learning requirements'.

Muggleton's successful trend of Machine Learning research opens us to some hope that teaching human-based creation of incompletely specified SRPS may benefit of the U-SL attitude, as the two following examples suggest.

Example 1, relative to symbiosis among the components of a system. In Section V.B, we have seen that symbiosis is better defined by its "vitally separation-sensitive interdependence" among the components, as illustrated by symbiotic Peano's primitive notions and axioms. As far as we know, teaching the recognition and handling of separation-sensitive interdependent systems, a skill necessary to creative programmers, does not exist yet. The research presented here provides a few clues of how it could be formalized. A tight collaboration with specialists in Cognitive Sciences should enable us to provide a large enough battery of symbiotic and non-symbiotic systems so that, mimicking US-L, we could unravel the deep features of systemic symbiosis, a necessary, if not sufficient, condition to safely handle creativity.

Example 2, relative to Oscillation. Oscillation has been introduced in Section V, where we underlined the difference between (P1) and (P2) problems. While exposing the "René's disciples" example, we used the switch between these two problems by replacing formula (10) by (11). Understanding the nature of a switch from a "$\forall\ \exists$" problem to a "$\exists\ \forall$" one is not easy, and its justification, as said above, requires epistemic considerations. We think that a strategy *à la* U-SL may constitute a tool favoring the understanding of the importance of the shift proposed here.

C. CSE and Computational Cognitive Science

Cartesian Systemic Emergence seems to us heavily related to the topic of human reasoning mechanisms, cognitive and computation models, human cognitive functions and their relationship, and even to modeling human multi-perception mechanisms. Moreover, scientific creation, as a particular human invention [22], becomes a highly economically interesting topic when it can be turned into an implementable science. CSE does try to build an implementable theory of SRPS scientific creation. In consequence, the four fundamental facets of CSE bring several stimulating challenges to Cognitive Science (CS). In this paper we would like to mention modular model of the brain [3] [31], frame problem [3] and conceptual blending [11].

Bermudez's work, as cited above, seems to imply that CS is somewhat wary of non-modular processing. One of the reasons is that non-modular processing very quickly meets frame problem-like difficulties. We have seen that, during Resonance Thinking, the human brain is rather at ease with the identifications needed to handle the frame-problem. Why

it is so? Is it because there is a particular kind of internal representation the human mind is able to construct? Alternately, is it because our mind includes mechanisms that are presently out of the scope of the current modular approach to our mind architecture [3]? Moreover, performing CSE includes a symbiosis of form, a meaning, a representation formalism, mechanisms and, importantly, includes also reaching a human agreement via conceptual coherence [35] and real-world exploitation. Does it mean that a modular approach to mind architecture should be revised? Could it be possible that some kind of symbiotic approach might be better suited even though it is more complex?

In [3], Bermúdez pointed out the influence of Computer Science on the development of CS Paradigms. CSE, as an example of symbiotic thinking (i.e., simultaneously focusing mentally on several different topics), represents a way of thinking, which - as far as we know - is not studied in CS. One cause may be that symbiotic thinking is considered as not achievable in CS. For instance, John Medina claims in [31] that our brain *is not conceived* to handle simultaneously several different topics. We may agree that it may be impossible for a non-trained person to perform two different physically challenging tasks. We believe, however, that this opinion, when generalized to mental processes, is born from existing brain synergic models (that are thus non-symbiotic) as well as from some possible misinterpretations of external observations. In particular, the observation of symbiotic modules in action may meet difficulties with comprehending the emergence of a solution in an active performance. At least, its explication is bound to seem obscure and a clear (but inexact) presentation of its functioning tends to explain the modules roles (once their interaction is completed), as if they were independent of each other, i.e., using a synergic model. The problem of spotting symbiotic interaction, in itself, is therefore hard to tackle. This difficulty becomes obvious when psychoanalysis describes harmful relationships of the sick person with his/her self. A solution to the problems seems to become possible when, as suggested by famous psychoanalyst C. G. Jung, a symbiotic solution starts to be built following the rule that: "… it is as much a vital necessity for the unconscious to be joined to the conscious as it is for the latter not to lose contact with the unconscious." ([25], section 457, p. 298). We could use a similar way of speech to express the fact that two modules of an emerging system should not 'lose contact' one with the other.

In complement to considering symbiosis as a reasonable paradigm in CS, we have tried to find some concepts of CS that resonate with CSE-thinking. We have found some similarities between Resonance Thinking and Conceptual Blending (CB) as presented in [11]. On a high-level of abstraction, RT and CB seem similar, since they are both concerned with the construction of meaning and they both involve 'merging'. Of course, they also show some differences at this high-level because RT is consciously performed, while CB is considered as taking place outside consciousness and is not available to introspection (as in [11], p. 33). We believe that this unconscious feature of CB disappears if people work in domains where rigor, justification and reproducibility of results are essential. Incidentally, let us point out that Fauconnier and Turner's illustrations, in [11], do not fulfill these stipulations.

At a lower-level, RT seems to us more complex than CB. Let us mention several features of RT that contrast CB.

- CB is highly nondeterministic while, in RT, the solution is specified in advance, even-though informally. Thus, RT performs what could be called a 'goal-oriented symbiosis'. While handling the generated experiments, RT focuses on what resonates as contributing to a universal solution, as in René's example, 'works' in (10).
- RT involves solving underspecified constraints due to the presence of incompleteness and an informal specification.
- RT not only handles a given data input (experiments) but it also generates complementary data (temporary facts, feedback, new experiments).
- CB is performed on mental spaces, i.e., small conceptual pockets constructed for purposes of local understanding and action ([11], p. 40). In RT, there are no small conceptual packets since global understanding is required even in considerations that may seem local.
- CB usually works with two mental spaces. RT, via topological symbiosis, works with three inputs (two experiments and one goal) and the solutions obtained are temporary until other experiments confirm the output.
- Fauconnier and Turner [11], in relation to CB, claim that researchers are unaware of how they are thinking. RT is a description of our way of thinking relevant to the creation of SRPS.
- In the case of CB, the effects of some unconscious imaginative work are captured by consciousness, but the operations that produce it are not ([11], p. 58). As said above, RT (and CSE) is a description of our way of thinking that is relevant to SRPS creation. This means that we are consciously aware of the informal specifications of the operations performed by our mind.

It follows that CSE might well be part of a challenge for CS. This will be achieved by developing CS models that capture all the essential characteristics of CSE, by finding methods and tools to study the emergence process in an active performance and developing on-purpose computational models for this particular way of thinking. Even though the topic is challenging, we are convinced that a strong desire or need to solve problems that CSE suggests to CS will lead soon or later to a fruitful empowerment of CS. We hope that the models presented in the present paper might be of help in such a difficult task.

We are aware that our description of the cognitive tasks involved in CSE does not provide a clear idea of whether it is possible (or reasonable) to find a way to break down the

cognitive tasks that are performed into more determinate tasks. We describe what humans do or what they have to do without specifying how these tasks are performed by our brain. We thus believe that research on these topics in the field of CSE in particular and its comparison with scientific creativity in general (i.e., a comparison with scientists' creative thinking in several scientific domains) might bring new conceptual and procedural switches not only in Computer and Cognitive Sciences, but also in other human activities.

D. Comparison with works on General System Theory

The next type of related work concerns complex systems modeling [4] [29] [30]. This falls to the domain of General System Theory (GST), where by a General System is understood "the representation of an active phenomenon perceived as identifiable by its projects in an active environment, in which it functions and it transforms teleologically" [30], p. 40. Similarly to our use of paradigm (P2), GST is conceived in the logics of an open unique global system. However, while GST responds actively to the problem of modeling and *observing* real-world phenomena, CSE builds a unified 'theory' of human *creation* of particular complex systems. It might, therefore, be seen as a kind of meta-theory of General System Theory used not in its standard observation mode but in its new 'creation' mode. Moreover, while, in GST, the conjunctive logic (i.e., conjunctive modular composition) is applied to already existing entities (i.e., the objects that exist already in their independent form; see [30], p. 33-41), in our work, symbiotic composition concerns 'objects' that start to exist only once the process of symbiotic composition is achieved.

E. Comparison with works on Multidisciplinary Systems Design

The last type of related work considered here concerns the systems design in general. It is thus somewhat related to the field called System Engineering (SEng) and Whole Systems Design, even though these fields do not use the same language and representations and the latter studies social and economical systems oriented towards sustainable solutions rather than developing problem-solving systems in general. The careful study of [7] [28], and [39] shows that SEng, even for complex systems, relies on modular compositions. However, even a complex communication of system modules, or their synergy (as in Whole Systems Design [5] [6]) is not a symbiosis. In order to have a symbiotic composition of the parts of the system it is necessary that these parts are defined in terms of the other parts of the system or even in terms of the system. Moreover these approaches work with specifications that are not informal and they do not consider prevention and control in the sense understood in our work. Therefore, our work on CSE and ST differs from these approaches to systems design. Note that CSE does not improve these approaches but it *enlarges* the class of possible approaches to Multidisciplinary Systems Design. Each of the mentioned approaches, by their competitive advantages, plays an important role for the progress in the field of Multidisciplinary Systems Design.

VII. CONCLUSION

The goal of Cartesian Systemic Emergence is to develop an implementable systems design theory for Symbiotic Recursive Pulsative Systems. In this paper, we have introduced one of its symbiotic features, namely Resonance Thinking. RT takes care of generating and handling experiments during the creation process of symbiotic systems specified, at the start, by an informal specification. RT is very complex, since it has to deal with the requirements of control and prevention, as well as with the process of 'shrinking' the incompleteness in accordance with the pulsation model.

Presently, our goal is not to apprehend all the conscious details of the operations performed by RT and CSE. Our present goal is to specify what enters into the 'game' of RT (and CSE) and what the 'winning strategies' are in order to conceive all the rules of 'the full game' of CSE. In other words, presently, we aim to develop a 'prosthesis' that can be implemented and used during CSE. We are convinced that apprehending human operations first by relevant informal specifications is halfway to a reasonable implementable solution. Thus, we believe that, even in its presently incomplete version, CSE brings forward thinking mechanisms that are essential for exploration, creation of possibilities, anticipation, resonance, blending, on-purpose creating of informally specified tools, invention, discovery, and so on.

Finally, recall that it is largely accepted that inspiration seems to take place anytime, such as while walking (e.g., Poincaré's case [36]), showering or during a pause playing the violin (e.g., Einstein's case). It is usually also accepted that some sort of unconscious incubation precedes this inspiration. Since we differentiate 'unconscious' and 'non-verbal', such an incubation does not take place during RT. Furthermore, contrary to Popper's opinion [37] that "there is no such a thing as a logical method of having new ideas, or a logical reconstruction of this process," RT is a systemic method for generating new and relevant ideas. Of course, its 'logical reconstruction' is not trivial, as is illustrated by this paper. CSE, with its four symbiotic facets, nevertheless seems to be a good start for a 'Cartesian reconstruction' of this process.

REFERENCES

[1] Y. Kodratoff and M. Franova, Resonance Thinking and Inductive Machine Learning, in S. Sendra Compte (eds.), Proc. of The Fourteenth International Conference on Systems, ICONS 2019, ISBN: 978-1-61208-696-5, 2019, pp. 7-13.

[2] F. Bacon, The Great Instauration, Complete Works of Francis Bacon, Kindle Format, Minerva Classics, 2013.

[3] J. L. Bermúdez, Cognitive Science: An Introduction to the Science of the Mind, Cambridge University Press, 2014.

[4] L. von Bertalanffy, General Systems Theory, George Braziller, 1969.

[5] J. L. Blizzard and L. Klotz, A framework for sustainable whole systems design, Design Studies 33(5), 2012, pp. 456–479.

[6] F. Charnley and M. Lemon, Exploring the process of whole system design, Design Studies 32(2), 2011, pp. 156-179.

[7] J. A. Crowder, Multidisciplinary Systems Engineering: Architecting the Design Process, Springer, 2018.

[8] R. Descartes, Discourse on the method (Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences), in R. Descartes, Œuvres philosophiques (3 vol.). Edition de F. Alquié. T. 1, Classiques Garnier, Bordas, 1988, pp. 567-650.

[9] R. Descartes, Principles of Philosophy (Les principes de la philosophie), in R. Descartes, Œuvres philosophiques (3 vol.). Edition de F. Alquié. T. 3, Classiques Garnier, Bordas, 1989, pp. 87-525.

[10] R. Descartes, Principles of Philosophy, in R. Descartes, translated by J. Cottingham, R.Stoothoff, D. Murdoch: Philosophical Writings of Descartes, vol. 1, Cambridge University Press, 2006, pp. 177-292.

[11] G. Fauconnier and M. Turner, The Way We Think: Conceptual Blending And The Mind's Hidden Complexities, Basic Books, 2003.

[12] V. Filkorn, Character of the contemporary science and its methods (Povaha súcasnej vedy a jej metódy), Vydav. Slovenskej Akadémie Vied, 1998.

[13] M. Franova, A construction of a definition recursive with respect to the second variable for the Ackermann's function, Rap. de Recherche No.1511, L.R.I., Univ. de Paris-Sud, Orsay, France, 2009.

[14] M. Franova, Cartesian versus Newtonian Paradigms for Recursive Program Synthesis, International Journal on Advances in Systems and Measurements, vol. 7, no 3&4, 2014, pp. 209-222.

[15] M. Franova and Y. Kodratoff, A Model of Pulsation for Evolutive Formalizing Incomplete Intelligent Systems, in L. van Moergestel, G. Goncalves, S. Kim, C. Leon, (eds): INTELLI 2017, The Sixth International Conference on Intelligent Systems and Applications, ISBN: 978-1-61208-576-0, 2017, pp. 1 - 6.

[16] M. Franova and Y. Kodratoff, Cartesian Systemic Emergence - Tackling Underspecified Notions in Incomplete Domains, in O. Chernavskaya, K. Miwa (eds.), Proc. of COGNITIVE 2018: The Tenth International Conference on Advanced Cognitive Technologies and Applications, ISBN: 978-1-61208-609-5, 2018, pp. 1-6.

[17] M. Franova, Trace of computation for ack(3,2), https://sites.google.com/site/martafranovacnrs/trace-of-computation-for-ack-3-2, retrieved 05/14/2020.

[18] M. Franova and Y. Kodratoff, Cartesian Systemic Pulsation – A Model for Evolutive Improvement of Incomplete Symbiotic Recursive Systems, International Journal On Advances in Intelligent Systems, vol 11, no 1&2, 2018, pp. 35-45.

[19] J. Gatial and M. Hejný, Construction of planimetry (Stavba planimetrie), Slovenske Pedagog. Nakladatelstvo, Bratislava, 1973.

[20] J. Y. Girard, Domain of the sign or the bankruptcy of reductionism, (Le champ du signe ou la faillite du réductionnisme), in T. Marchaisse, (dir.): Le théorème de Gödel, Seuil, 1989, pp. 145-171.

[21] K. Gödel, Some metamathematical results on completeness and consistency, On formally undecidable propositions of Principia Mathematica and related systems I, and On completeness and consistency, in J. van Heijenoort, From Frege to Godel, A source book in mathematical logic, 1879-1931, Harvard University Press, Cambridge, Massachusets, 1967, pp. 592-618.

[22] J. Hadamard, An Essay on the Psychology of Invention in the Mathematical Field, Read Books, 2007.

[23] S. Hawking, The Theory of Everything: The Origin and Fate of the Universe, Jaico Publishing House, 2008.

[24] M. Hejny, The roots of causal thinking and Greek mathematicians (Korene kauzálneho myslenia a grécki matematici), in S. Znam, L. Bukovsky, M. Hejny, J. Hvorecky, B. Riecan: Pohlad do dejín matematiky, ALFA-SNTL, 1986, pp. 11-83.

[25] C. G. Jung, Symbols of transformation, Princenton University Press, 1956.

[26] D. Kahneman, Thinking, Fast and Slow, Allen Lane, 2011.

[27] F. Bacon, The Advancement of Learning, Kessinger Publishing Company, 1994.

[28] H. Kopetz, Simplicity Is Complex: Foundations of Cyber-physical System Design, Springer, 2019.

[29] J. L. Le Moigne, Theory of the general system (La théorie du système général, théorie de la modélisation), P.U.F, 1984.

[30] J. L. Le Moigne, Complex systems modeling (La modélisation des systèmes complexes), Dunod, 1999.

[31] J. Medina, Brain Rules, Pear Press, 2008.

[32] D. Michie, Machine learning in the next five years, Proceedings of the third European working session on learning, Pitman, 1988, pp. 107-122.

[33] S. Muggleton, D. Lin, and A. Tamaddoni-Nezhad, Meta-interpretive learning of higher-order dyadic datalog: predicate invention revisited, Machine Learning 100; 2015, pp. 49-73.

[34] S. Muggleton, U. Schmid, C. Zeller, A. Tamaddoni-Nezhad, and T. Besold, Ultra-Strong Machine Learning: comprehensibility of programs learned with ILP, Machine Learning 107, 2018, pp. 1119-1140.

[35] G. L. Murphy and D.L. Medin, The Role Theories in Conceptual Coherence, Psychological Review 92(3), 1985, pp. 289-316.

[36] H. Poincaré, Invention in Mathematics (L'invention mathématique), in J. Hadamard, Essai sur la psychologie de l'invention dans le domaine mathématique, Editions Jacques Gabay, 1993, pp. 139-151.

[37] K. Popper, The logic of scientific discovery, Harper, 1968.

[38] V. K. Ranjith, Business Models and Competitive Advantage, Procedia Economics and Finance 37, Elsevier, 2016, pp. 203 – 207.

[39] C. S. Wasson, System Engineering Analysis, Design, and Development: Concepts, Principles, and Practices, Wiley-Blackwell, 2015.

[40] A. Yasuhara, Recursive Function Theory and Logic, Academic Press, New York, 1971.

[41] H.P. Zwirn, Complex systems : Mathematics and biology, (Les systèmes complexes : Mathématiques et biologie), Odile Jacob, 2006.

Assessing the System Condition Based upon Limited Maintenance Data of the Taipei Metro System and Estimating its Remaining Lifetime

Tzu-Chia Kao and Snow H. Tseng
Department of Electrical Engineering,
National Taiwan University,
#1, Sec. 4, Roosevelt Rd, Taipei 10617, Taiwan
e-mails: b03901004@ntu.edu.tw, stseng@ntu.edu.tw

Abstract—**In this study, we analyze the maintenance data provided by Taipei Rapid Transit Corporation; the data consists of recent Taipei metro system maintenance record. However, it is unclear whether the lifetime information of the system is contained within the partial-lifespan maintenance record of a still-young metro system. Furthermore, the Taipei metro system is renowned for its performance and well maintenance, which further complicates the analysis. Based on limited maintenance records, the research objective is to explore the feasibility of extracting reliable information that is indicative of the current stage of life and the remaining lifetime of the metro system.**

Keywords-degradation; maintenance; metro; MRT; performance analysis; data analysis.

I. INTRODUCTION

In line of the recent reported research study [1], we further analyze the maintenance record of the Taipei Mass Rapid Transit (MRT) system, which is a well-maintained system renown worldwide for its tidiness, stability and efficiency. The Taipei metro system of Taipei Rapid Transit Corporation (TRTC) began operation on March 28, 1996; it has been operating for 24 years. Since the MRT TRTC is relatively young, most of the equipment has not yet been replaced; the records of repair and maintenance are detailed and are stored in digital form. The goal of this research is to determine whether it is possible to acquire status information of the system from the limited time-span maintenance records. Such information may be crucial for improving the performance of the Taipei MRT system. More generally, this research may shed light on enhancing performance of other metropolitan MRT systems.

Here we investigate the feasibility of extracting information from the Taipei MRT maintenance record and determine whether it is possible to acquire reliable information of the system performance from the limited time-span maintenance records. If the maintenance data indeed contains such information of the current system status, our goal is to assess the current stage of life of the Taipei MRT system and determine the remaining lifetime.

The performance and degradation of metropolitan metro systems have generally been the focus of the daily public. Various studies have been reported, including technical issues of the MRT. Rail track condition monitoring is an important technical concern of the MRT system [2]. However, constant monitoring of the MRT system is not available, typically the usual maintenance is performed once a month or less. Track condition has attracted much attention since it is a potential threat to the railway system. Studies to prevent such problem have been reported [3]-[5]. The general goal of the research and technical modifications is to improve the performance and reliability of a mass rapid transit system.

Comparing information of other metropolitan mass transport systems may be helpful. As reported in [6], train R36 of the New York metropolitan subway serviced from 1964 to 2003, a total of 39 years. R160s were used to replace 45-year-old trains. News reported that the oldest trains of the New York City Subway were planned to serve for 58 years, and later this type of train were actually found too old with very high failure rate and not appealing to the general public [7]. The subway train lifetime was estimated to be around 40 to 50 years. For example, since 1987, some lines of Singapore Mass Rapid Transit (SMRT) have been operating for 30 years. Thus, the actual wear-out period of a metro system, assuming they are similar, may roughly lie between 20 years (the oldest TRTC asset), and 40 years (New York City Subway). However, all of these metropolitan metro systems are different in various aspects, including: model, company, maintenance, management culture, etc. It is natural that the characteristics of these MRT systems are not the same or even far from similar. With multiple factors entangled, the problem to accurately assess the system performance and age may be very challenging.

Assessment and quantification of the system current status are essential to improving performance; to accurately assess the system current status is often nontrivial. The degradation curve is commonly employed for estimation of the system current status. Analysis of the reliability relies on the failure rate and maintenance records [8]. Based on the status of the system, possible improvement of the

maintenance and performance can be assessed. To study the maintenance and performance characteristics, various approaches have been reported [9]-[17], including the popular bathtub curve analysis [18]-[23]. Typically the Bathtub-shaped curve is employed for system performance analysis [24]. Analysis based upon the bathtub curve has been extensively applied to various problems; modifications to improve applicability have been reported [11], [25], [26]. It is possible that the bathtub-shaped curve could be affected by human factors; for example, if the asset retired in its early stage, the curve may not rise up within the wear-out period and may even descend. If properly maintained, the curve may not rise in the wear-out period, similar to the situation in airline industry. However, few MRT systems in reality exhibit degradation behavior similar to the bathtub-shaped curve model [27]. It is possible such bathtub curve may not be the ideal model for analyzing the metro system performance.

The rest of this paper is organized as follows: Section II describes the goal of this research project. Section III describes the research method. Section IV summarizes the data analysis. A summary is presented in Section V, a conclusion in Section VI, and lastly, a description of the Future Work in Section VII, followed by an acknowledgement.

II. Research Goal

We propose a thorough investigation of the Taipei MRT data. Available data consists of 11 MRT systems: electric multiple unit (EMU) propulsion, EMU air conditioner, EMU communication, switcher, platform door, 22kV switchboard, automated fare collection (AFC) door, Wenhu Line traffic control computer, transmission system, elevator, and escalator. By analyzing the dataset with various approaches, such as deep learning, the research objective is to analyze the current status of the system, and identify possible tendencies or features that may be indicative of the system performance.

III. Methods

The bathtub-shaped curve model [24] is commonly employed to assess the system condition. It consists of a break-in trend as the system condition improves, followed by a plateau regime where the system condition is stable; after the stable regime, the system condition withers with increased malfunction rate, followed by a steep increase of malfunction rate as the system breaks down. Together, the bathtub-shaped curve represents the various stages of an ideal system.

The bathtub-shaped degradation curve is a theoretical model used in many problems. It is an idealized evolution trend of a system. The feasibility of applying such bathtub-shaped curve may depend on the application. Specifically, the system condition may not follow the same degradation curve, also, each equipment system may exhibit different characteristics depending on the specific application. Furthermore, each equipment in the Taipei metro system consists of various brands and various models that may possess different intrinsic characteristics. Together, all these factors intertwined complicates the situation and it is very challenging to dis-entangle the complex problem.

Ideally, analysis of the maintenance data would yield a simple bathtub-shaped degradation curve for each equipment. However, most systems do not follow the same degradation curve, not to mention each equipment may exhibit different characteristics. Since each equipment is maintained by human, the degradation curve is unlikely to be a simple universal bathtub-shaped curve. By analyzing the maintenance data, our goal is to decipher the feasibility to assess the MRT current stage of life based on available data ranging over a limited time-span.

IV. Data Analysis

We use the maintenance records provided by TRTC. The AFC gates statistics are shown in Figure 1; the average malfunction rate is calculated and shown in Figure 2. As shown in Figure 1, the equipment consists of mixed brand, model, age; thus, the total number of malfunctions is not representative of individual equipment.

Based on the data analysis, it is found that the failure rate has been declining each year. It is speculated that this declination is due to the TRTC's maintenance becoming better and better. On the other hand, because of the difference on the age and amount of the equipment, the total number of malfunctions is not a direct indication of the failure rate of a specific equipment.

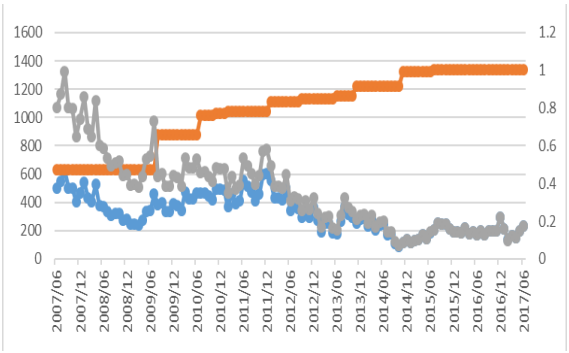


Figure 1. The malfunction rate of AFC gates vs. time. The failure rate decreases monotonically with time. (Blue line): the total number of incidences per month; (orange line): the total number of equipment (as the MRT system is expands, the total number of equipment increases; (gray line): the monthly malfunction rate.

With sufficient maintenance data of escalators and elevators, we try to quantify the failure rate *vs.* the age of each individual equipment. On the other hand, for some equipment that seem to have no difference among each other, e.g., EMU, platform door, switcher and 22kV switchboard, we use the original method of analyzing the failure times against failure date. As for the Wenhu line Central control

computers, AFC gates and transmission system, we cannot decipher the replaced new ones from others, so we employ the latter method.

The following data of the number of escalators, elevators, equipment of individual station and the operation time of individual equipment are acquired from the Internet data and the records given by TRTC, and may differ from the actual number.

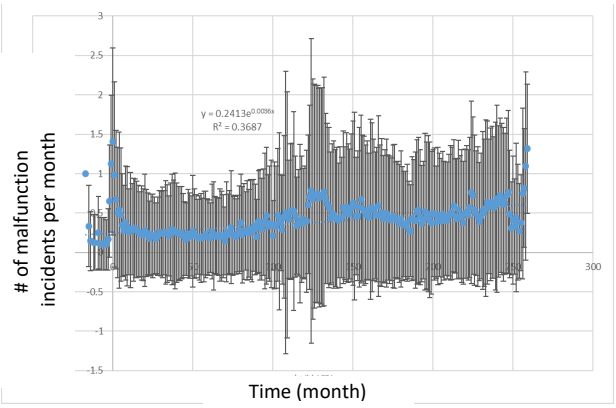


Figure 2. The malfunction rate as a function of time.

Notice that the malfunction rate of Figure 2 rises on both ends of the curve. The malfunction rate increases with time as the system becomes frail and old. Also, the data points on the left and right side of the curve each belongs to different groups of systems. It is speculated that the systems of the left side may have received adequate maintenance from the early years and therefore exhibits a lower malfunction rate; whereas the systems of the right side received adequate maintenance in its later years and therefore exhibits a higher malfunction rate.

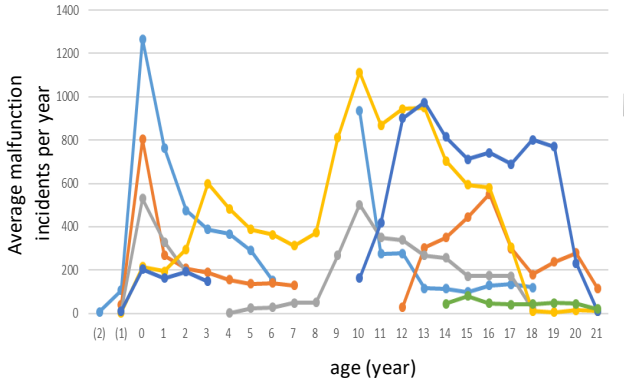


Figure 3. The average malfunction incidents each year for different subway lines.

We also tried categorizing the maintenance record according to different subway lines (Figure 3). Notice that the data span varies for each subway line, since some subway stations were built much earlier and some were built later. However, there are possible pitfalls with such categorization: each subway line has different number of stations; even along the same subway line, the age of each station also differs. For example, for the Songshan-Xintien Line, its stations roughly can be divided into two groups: stations that have been established for more than ten years, or less than ten years. If each sub-group has a characteristic peak, then the overall malfunction rate of Songshan-Xintien Line may exhibit two peaks due to each sub-groups. Analyzing the same escalator all employed the same time should be more appropriate.

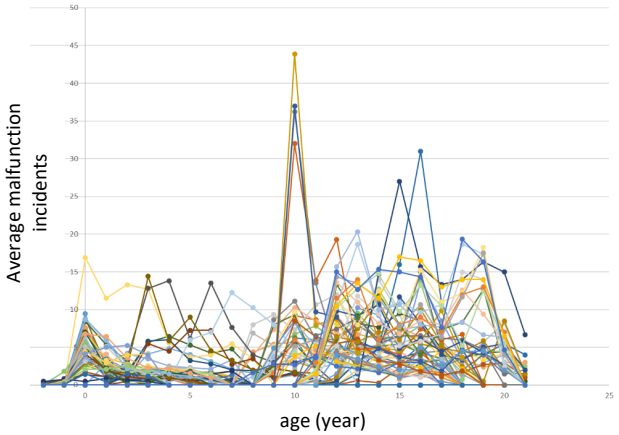


Figure 4. The average malfunction incidents per month as a function of station age. Each curve is acquired by: the total number of escalator malfunction incidents divided by the total number of escalators in each station.

As shown in Figure 4, the malfunction incidents as a function of station age is sporadic. It is infeasible to add up all these curves of different stations, because some stations have more escalators and the weighting differs. Nevertheless, though sporadic, the average malfunction incident rate is clearly lower for younger stations.

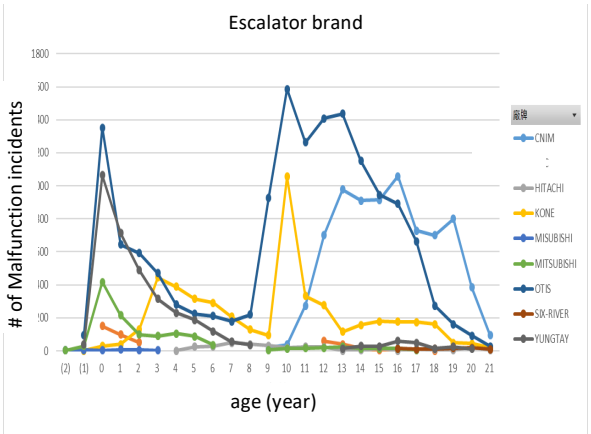


Figure 5. The average malfunction incidents per escalator vs. age (year) for different brands. The trend varies for different brands.

In Figure 5, the trend varies with different brands. In addition, the number of escalators of each brand is not the same and therefore the comparison may not be conclusive.

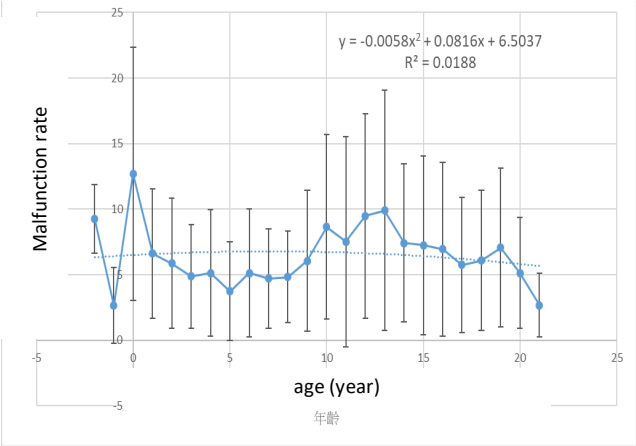


Figure 6. The elevator average malfunction rate vs. age (year).

Figure 6 is obtained by averaging the malfunction rate of each elevator. There is some increase followed by decrease in the middle. Overall, there is no clear trend of increase or decrease.

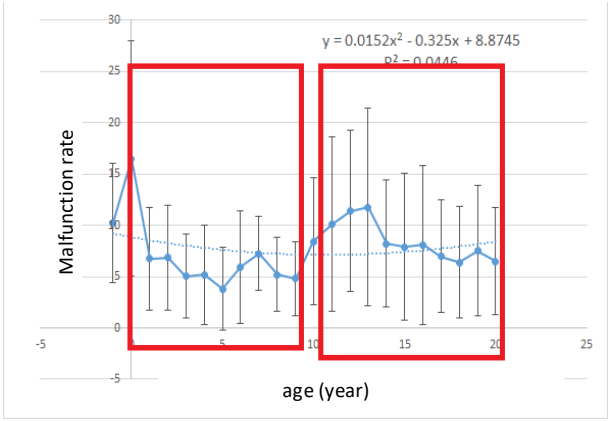


Figure 7. The elevator average malfunction rate vs. age (year).

The elevator average malfunction rate vs. age (year) is shown in Figure 7. The data was reorganized; the trend as shown can be separated into two groups (the two red boxes as indicated in Figure 7). It seems that the younger group of elevators (left red box) and older group (right red box) of elevators both exhibit a decrease in the malfunction rate with time. It is suggested that the maintenance technique may have improved or maintenance has been employed more frequently, so that the malfunction rate degrease with time, regardless of the specific age of the elevator.

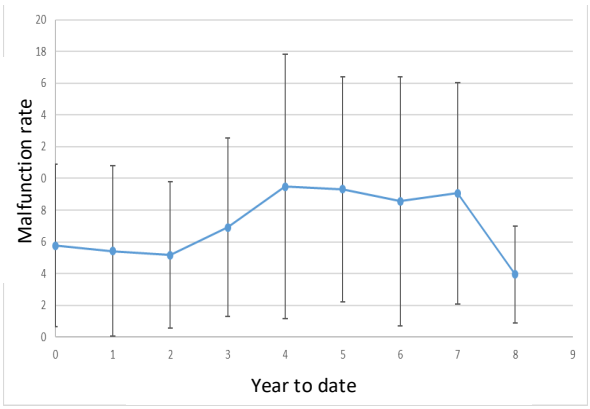


Figure 8. The average elevator malfunction rate vs. number of years to date.

Figure 8 is the malfunction rate plotted against the number of years to date. The overall variation does not show a monotonic trend. It is suggested that the malfunction rate of the elevator is not directly dependent upon the elevator age.

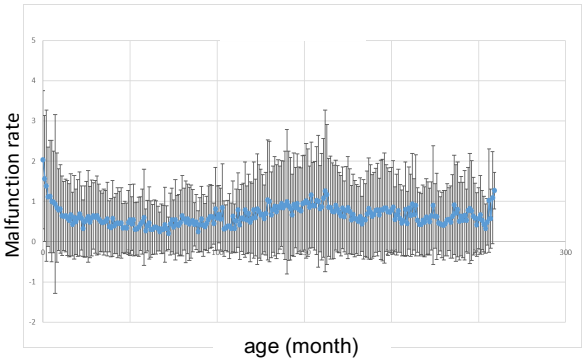


Figure 9. The escalator average malfunction rate vs. age (year).

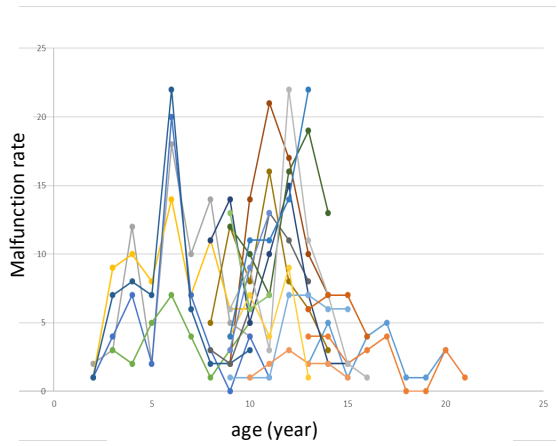


Figure 10. The malfunction rate as a function of age for 20 elevators. The variation is large and sporadic.

The elevator average malfunction incidents vs. age (year) is depicted in Figure 9; the trend is similar to Figure 8. The variation between each month is not pronounced; in addition, increase, decrease, occurs sporadically. Overall, the characteristics are not apparent.

Figure 10 is the malfunction rate as a function of age for 20 elevators. The plotted data exhibits large variations with sporadic behavior. It is difficult to come up with an average trend to represent the general behavior of these elevators.

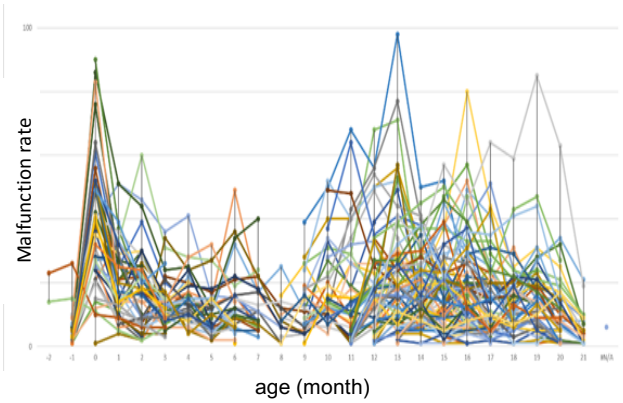


Figure 11. the number of elevator malfunction incidents vs. age.

The number of elevator malfunction incidents as a function of age is shown in Figure 11. As compared to the escalator, there is not apparent increase of the malfunction rate similar to the elevator case (Figure 2). Then we analyze the maintenance record of the MRT transmission system. Though the total number of transmission system is given, the age of each transmission system cannot be deciphered.

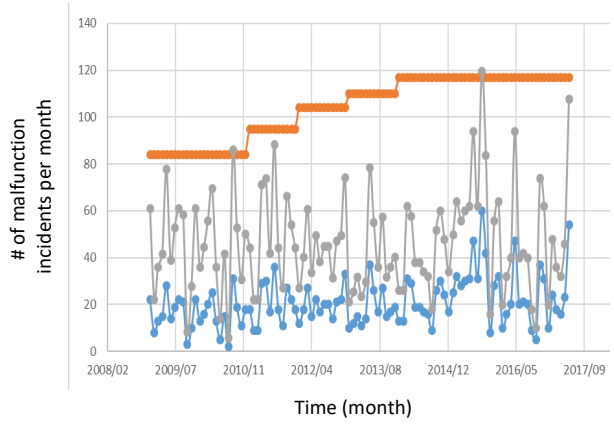


Figure 12. The total malfunction incidents per month as a function of time. (Blue line): the total malfunction incidents; (orange line): the total number of transmission systems; (gray line): the ratio of malfunction incidents over the total number of transmission systems.

As shown in Figure 12, the total number of transmission systems increases with time, and the number of malfunction incidents also increased slightly with time. The malfunction ratio (gray line) does not show an overall increase with time.

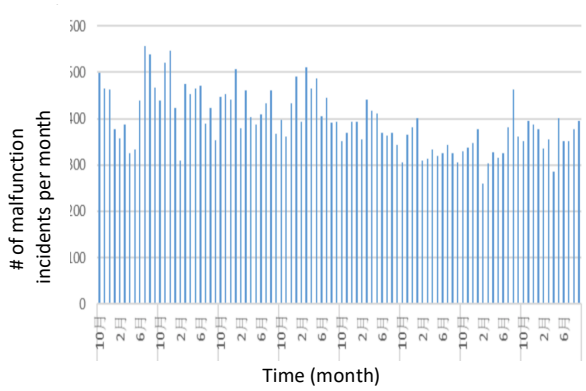


Figure 13. The total number malfunction counts of the escalators per month.

As shown in Figure 13, the total number of malfunction counts per month appears to be steady. However, since the total number of equipment varied with time, this steadiness of the escalator malfunction rate is not conclusive. More data is required to ascertain its general trend.

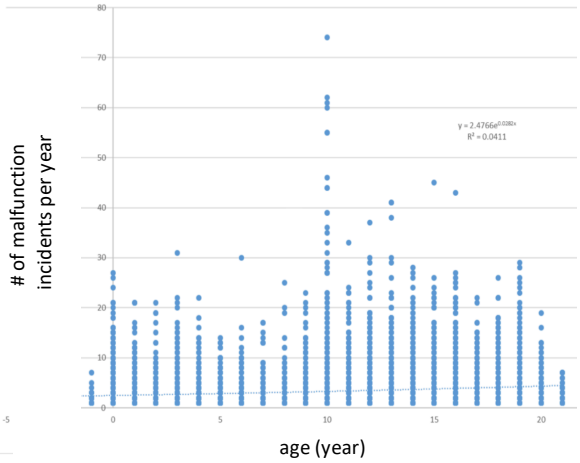


Figure 14. The number malfunction incidents of individual escalator per year as a function of age. The malfunction seemed to increase slightly with age.

The malfunction count per year for an individual escalator is plotted in Figure 14. The decrease near age of 20 years may be due to incomplete data. Overall, the malfunction count per year is steady and gradually rises with time.

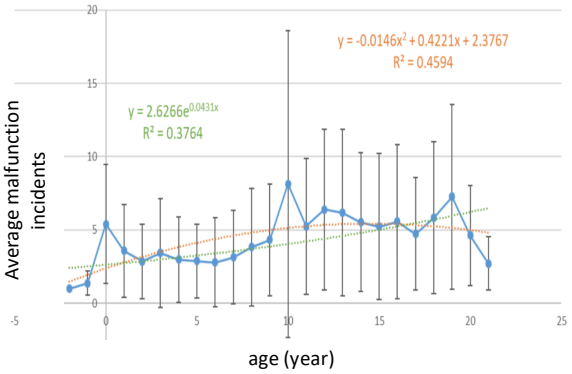


Figure 15. The average escalator malfunction rate. This is the average malfunction rate acquired from Figure 14. Notice that the overall malfunction rate gradually increases with age.

Figure 15 is the average escalator malfunction rate, showing a trend of gradual increase with the escalator age. However, the overall malfunction rate should increase with age; we anticipate the general trend increases if the data range is expended to 40 years or more.

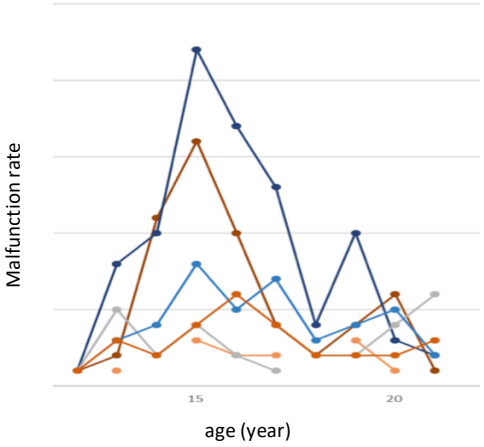


Figure 16. Malfunction rate of individual escalators vs. age. It is shown that the individual variation is pronounced; the relationship of individual malfunction rate and age is not decisive.

Figure 16 is the malfunction rate for six individual escalators. Notice that the individual variation is pronounced, suggesting that malfunction rate is not solely dependent upon the age of the escalator.

Trend of the maintenance data showed that the failure rate has been declining every year until it reached low and stable end tail. Possible factors are analyzed; this mostly likely is due to the improvement of MRT maintenance. On the other hand, since the age of each equipment and the number of samples for each equipment are not consistent, the total failure rate is not a fair representation of a specific individual equipment. Thus, our goal is to analyze the age's effect of the individual equipment.

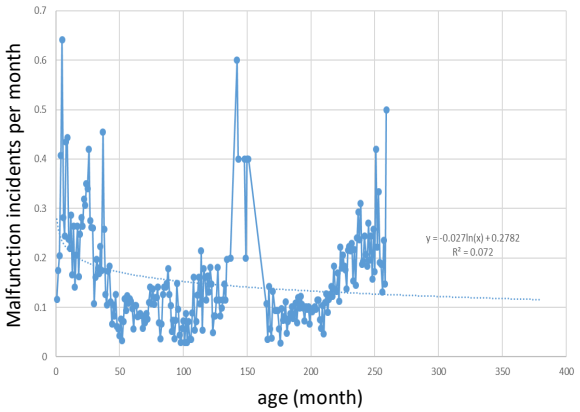


Figure 17. The average malfunction rate of the AFC gate *vs*. time.

With sufficient maintenance data of escalators and elevators, we quantify the failure rate vs. the age for each individual equipment with correlation analysis. On the other hand, for some equipment that seem to have no variation among one another, e.g., EMU, platform door, switcher and 22kV switchboard, we propose to calculate the malfunction rate vs. date of incident. As shown in Figure 17, the Wenhu line Central AFC gate failure rate vs. date is analyzed. The spikes at the center is likely due to lack of sufficient data points to reveal detail trend. Overall, the malfunction rate decreases and later increases, exhibiting behavior similar to a bathtub curve.

Then we analyze the maintenance data of the escalator system, elevator system, and EMU air conditioner of the metro system. These three systems are essential components of the metro system. The elevator and escalator are used regularly by the commuting people.

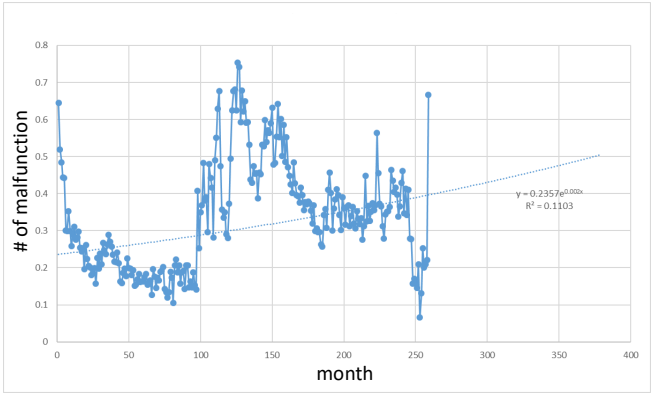


Figure 18. The reported average number of malfunctions each month of the Taipei metro station escalators.

There are 1144 escalators that have been recorded. From Figure 18, it is apparent that the malfunction rate increases with time. Nevertheless, there are two major age groups of escalators; it is unclear the trend of the younger group and older group of escalators are the same. Data analysis suggests

that these two group should be analyzed separately instead of as a whole.

The maintenance records of the elevator, escalator, and EMU air conditioner are analyzed. As shown in Figure 18, the maintenance record of the escalator is depicted; trend of the malfunction rate is irregular. Maybe due to limited span of the maintenance data, the specific stage of life-time is not apparent; further analysis is required.

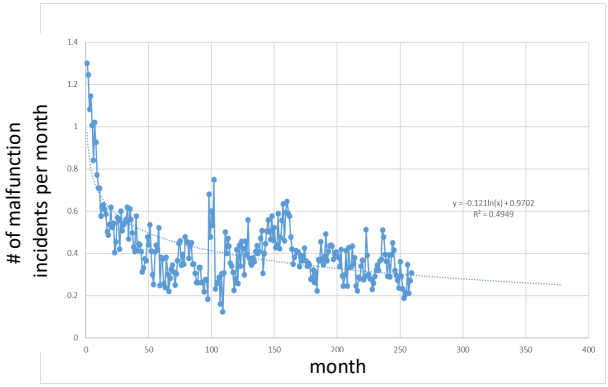


Figure 19. The reported average number of malfunctions each month of the Taipei metro station elevators.

Next, the maintenance record of the elevator system is shown in Figure 19. The maintenance record decreases with time monotonically. This monotonic trend appears to be a good match with the ideal bathtub-shaped curve.

Lastly, the maintenance record of the EMU air conditioner is shown in Figure 20, also exhibiting a decrease in its malfunction rate. However, compared to the smooth trend of Figure 8, the EMU air conditioner maintenance data is more volatile. Though the maintenance data of the three systems differ, they all roughly decrease with time, suggesting that the system is still young whereas the performance is still improving. Or, the system is being well-maintained such that the malfunction rate does not reflect the system performance accurately.

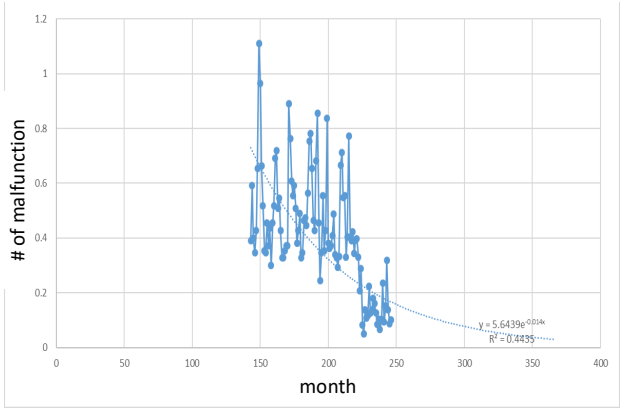


Figure 20. The reported average number of malfunctions each month of the Taipei metro EMU air conditioners.

The maintenance record provided by TRTC consists of only number of malfunction incidents per month. However, the severity of each malfunction may be drastically different, from as simple as the replacement of light bulb, up to power combustion resulting in system complete breakdown. Yet, in the available maintenance records, there is no information describing the severity of the malfunction event. Statistical analyses of these three systems (escalator, elevator, and EMU air conditioner) exhibit no apparent degradation of the system. Trends of the maintenance records suggest that the system condition improves with time, which is not the typical characteristics of a withering system.

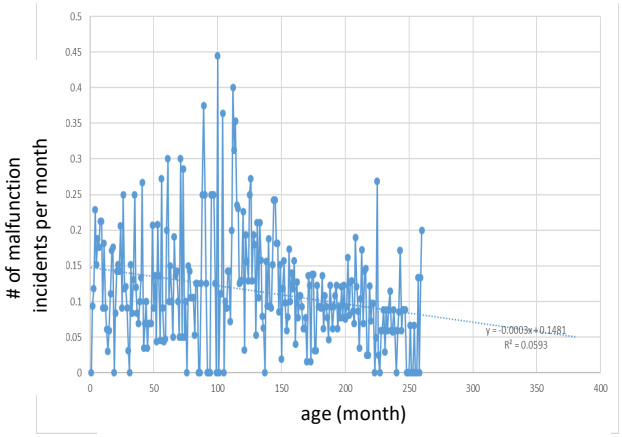


Figure 21. The average Taipei metro station elevators malfunction rate as a function of age (month). The general trend of the sporadic data decreased slightly.

More data point is required to determine the general trend of the malfunction rate. As shown in Figure 21, the trend decreases slightly with age; additional data point would be helpful to pinpoint the overall malfunction rate trend. Similarly, in Figure 22, more data is required to depict the trend of the malfunction rate of Muzha Line MRT platform screen door.

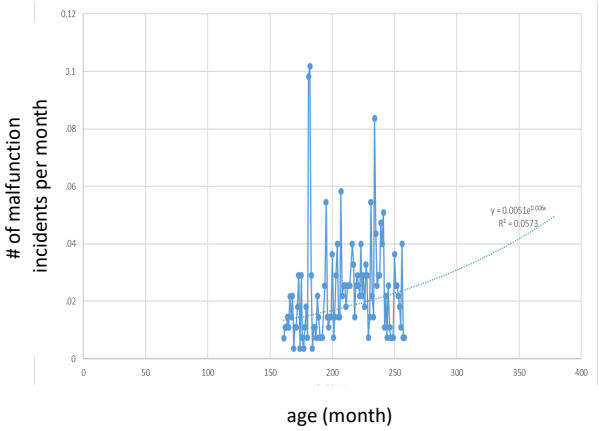


Figure 22. The average malfunction rate of the Muzha Line MRT platform screen door. The maintenance data showed a slight increase in malfunction rate with time.

V. DISCUSSION

Taipei Metro is still very young compared to other metropolitan metro systems in the world. For instance, the first subway route of New York City began operating in 1904 [6]; Tokyo Metro began operating in 1927. As mentioned in related data survey, New York Subway replaces its trains about every 40 years, far older than TRTC's age. Based on the available Taipei metro maintenance record, the objective of this research is to assess the current condition of the metro system, and furthermore, if possible, to estimate the remaining lifetime of each system.

By means of data analysis, our goal is to identify characteristics indicative of the current status of the metro system, remaining lifetime, and extrapolate its future trend. Data analysis of the Taipei metro system maintenance record revealed general trend and characteristics of the system condition and performance, however, information regarding the total lifespan of the system, current stage-of-life, and remaining lifetime have not been ascertained. Possible reasons include: i) the system is still in its early stage, ii) the regular maintenance altered the natural deterioration trend, iii) the maintenance records being insufficient both in quantity and variety. As a result, the estimation of system remaining lifetime based on the limited maintenance data is very challenging.

On the other hand, using the count of malfunction incidents reported each month to assess the performance of Taipei Metro system is not very sensitive or informative. There are various types of malfunction, small glitches in the system, or, as large as a system failure, each of these incident reports are treated and weighed equally as a single event. Thus, the recorded failure rate is not an accurate measure of the system performance. Furthermore, each equipment consists of various brands and models, it may be infeasible to statistically come up with a universal degradation relationship for such complex systems.

Furthermore, with essential components replaced by new ones through regular maintenance is like renewing the system lifetime, which further complicates the estimation of system lifetime. As for when an equipment is due to be replaced is determined by a wide range of factors, including: economic consideration, new technology, the public opinion for a trendy new system, etc. Oftentimes the equipment may be operating properly, yet, it is retired and replaced for non-technical reasons such as politics or public opinion, etc. Thus, for pragmatic concerns, lifespan of the equipment is not the critical factor for retirement/replacement.

On a more fundamental level, it is undetermined that the desired information can be extracted from the maintenance record. The specific information may be entangled and infeasible to be extracted from the limited maintenance records. More importantly, the desired information may not be fully contained within the dataset. The maintenance record may cover only a small fraction of the entire system lifespan. Therefore, the accuracy to estimate the system lifetime based upon the provided information may be limited.

Even though the malfunction rate may not be a sensitive and reliable measure of the Metro system present status, it can still provide useful information: i) the maintenance is considered appropriate if the failure rate continues to stay low; ii) the maintenance needs to be renovated if the failure rate continues to increase. We suggest recording more parameters that are indicative of the equipment status, such as: recording the number of passengers using the equipment, the operating time, the number of passengers effected by the failure incident, the cost of repair, or even the weather condition such as humidity or earthquake. Recent research [7] suggested that system performance can be better assessed with more information used to monitor the system; the analysis outcome can better reflect the system status and remaining lifetime.

Based on the available maintenance record of the Taipei metro system, statistical analysis indicates that no signs of deterioration or withering. The data analysis falls short to yield information regarding the total lifespan, the current stage of life, and the remaining lifetime. If data with longer span and more variables is available, it is possible that such information can be ascertained. On the other hand, data analysis shows that the maintenance of the Taipei metro system is adequate.

Overall, TRTC is doing a fine job in terms of maintenance and repair, which is clearly reflected in its low malfunction rate. Based on the maintenance record, it is reasonable to extrapolate and predict that the Taipei Metro system can maintain its current level of performance for a good 20 to 60 years without significant degradation. Since "trendiness" cannot be achieved via maintenance, if the Metro system is renovated every few years with new technology and trendy products such as slick new cable cars with new technologies; brighter flat displays with higher contrast, smart tint windows, etc., it is possible that Taipei Metro system can maintain a constant modern slick appearance that differs from other metropolitan metro systems (such as Tokyo, New York, Boston, etc.) that appears to be aged. As the maintenance data has shown, the Taipei Metro system is performing well and data analysis suggests that it would likely continue its performance at its current standard.

VI. CONCLUSIONS

The maintenance record provided by TRTC contains only number of malfunction incidents per month; the record contained no severity information or detailed description of the specific malfunction. Thus, the simple malfunction rate record falls short to provide an informative description of the status of the Taipei metro system, which involves enormous number of factors including human. Such complex system exhibits sophisticated characteristics that can hardly be characterized with a single variable. Thus, it is unrealistic to decipher the system status with just a single variable.

Nevertheless, we believe rich information is embedded within the maintenance records and can be harnessed with the appropriate analysis tools. Possibly with big data analysis, one can decipher more information that may help enhance the MRT performance. Instead of a single parameter (number of malfunction incidents per month), with more recorded data (i.e., longer timespan, wider variety of variables, and detailed description of each malfunction incident), combined with big data analysis, it would be more feasible to yield information indicative of the system condition to better monitor the system condition.

VII. FUTURE WORK

In order to extract information indicative of the system status, we propose the following:

First, based upon the original maintenance records, calculate the duration between: i) when the equipment was first engaged in operation, and ii) the failure date. This time interval represents the duration of malfunction-free operation, which is also the time for a malfunction to take place. By analyzing the malfunction-free duration instead of the number of malfunctions each month may yield more realistic relationship.

Second, calculate the failure times per month to acquire the failure rate for each individual equipment. (For the equipment related to EMU, the failure rate is acquired by the failure time divided by the total mileage.)

Third, average each equipment's failure times to get the average failure rate. Use a statistical software to calculate the regression curve, and extrapolate to compare with other metro system performance data.

The above three steps are to be performed repeatedly. Each step is employed to process another set of data. The difference between each dataset is compared and analyzed for its regression behavior and general trend. Based on experience analyzing the Taipei MRT dataset, we believe the above approach would enable extraction of crucial information that is indicative of the system status.

ACKNOWLEDGMENT

We thank TRTC for providing maintenance records for analysis and very helpful and responsive with all of our questions. This research is supported by the Taiwan National Science Council Grant MOST-106-2112-M-002-008 and MOST-107-2112-M-002-011.

REFERENCES

[1] S. H. Tseng and T.-C. Kao, "Investigating the Feasibility to Estimate System Performance Based Upon Limited Data of the Taipei Metro System," in *The Fourteenth International Conference on Systems*, Valencia, Spain, S. S. Compte, Ed., March 24-28, 2019, pp. 69-72.

[2] X. K. Wei, F. Liu, and L. M. Jia, "Urban rail track condition monitoring based on in-service vehicle acceleration measurements," *Measurement,* vol. 80, pp. 217-228, Feb 2016, doi: 10.1016/j.measurement.2015.11.033.

[3] M. Molodova, M. Oregui, A. Nunez, Z. L. Li, and R. Dollevoet, "Health condition monitoring of insulated joints based on axle box acceleration measurements," *Engineering Structures,* vol. 123, pp. 225-235, September 2016, doi: 10.1016/j.engstruct.2016.05.018.

[4] G. Lederman, S. H. Chen, J. Garrett, J. Kovacevic, H. Y. Noh, and J. Bielak, "Track-monitoring from the dynamic response of an operational train," *Mechanical Systems and Signal Processing,* vol. 87, pp. 1-16, Mar 2017, doi: 10.1016/j.ymssp.2016.06.041.

[5] R. Jiang *et al.*, "Network operation reliability in a Manhattan-like urban system with adaptive traffic lights," *Transportation Research Part C-Emerging Technologies,* vol. 69, pp. 527-547, August 2016, doi: 10.1016/j.trc.2016.01.006.

[6] Metropolitan Transportation Authority. "New York City Transit - History and Chronology." Retrieved Jan. 1st, 2018, from http://web.mta.info/nyct/facts/ffhist.htm

[7] D. Rivoli. "Ancient subway trains on C and J/Z lines won't be replaced until 2022, documents say." http://www.nydailynews.com/new-york/ancient-subway-trains-won-replaced-2022-article-1.2323289

[8] H. Yin, K. Wang, Y. Qin, Q. Hua, and Q. Jiang, "Reliability analysis of subway vehicles based on the data of operational failures," *EURASIP Journal on Wireless Communications and Networking,* journal article vol. 2017, no. 1, p. 212, December 15 2017, doi: 10.1186/s13638-017-0996-y.

[9] Z. G. Li, J. G. Zhou, and B. Y. Liu, "System Reliability Analysis Method Based on Fuzzy Probability," *International Journal of Fuzzy Systems,* vol. 19, no. 6, pp. 1759-1767, December 2017, doi: 10.1007/s40815-017-0363-5.

[10] A. Z. Afify, G. M. Cordeiro, N. S. Butt, E. M. M. Ortega, and A. K. Suzuki, "A new lifetime model with variable shapes for the hazard rate," *Brazilian Journal of Probability and Statistics,* vol. 31, no. 3, pp. 516-541, Aug 2017, doi: 10.1214/16-bjps322.

[11] T. Kamel, A. Limam, and C. Silvani, "Modeling the degradation of old subway galleries using a continuum approach," *Tunnelling and Underground Space Technology,* vol. 48, pp. 77-93, April 2015, doi: 10.1016/j.tust.2014.12.015.

[12] D. Brancherie and A. Ibrahimbegovic, "Novel anisotropic continuum-discrete damage model capable of representing localized failure of massive structures: Part I: theoretical formulation and numerical implementation," *Engineering Computations,* vol. 26, no. 1-2, pp. 100-127, 2009, doi: 10.1108/02644400910924825.

[13] R. Tahmasbi and S. Rezaei, "A two-parameter lifetime distribution with decreasing failure rate," *Computational Statistics & Data Analysis,* vol. 52, no. 8, pp. 3889-3901, April 2008, doi: 10.1016/j.csda.2007.12.002.

[14] C. F. Daganzo and N. Geroliminis, "An analytical approximation for the macroscopic fundamental diagram of urban traffic," *Transportation Research Part B-Methodological,* vol. 42, no. 9, pp. 771-781, Nov 2008, doi: 10.1016/j.trb.2008.06.008.

[15] C. Kus, "A new lifetime distribution," *Computational Statistics & Data Analysis,* vol. 51, no. 9, pp. 4497-4509, May 15 2007, doi: 10.1016/j.csda.2006.07.017.

[16] C. D. Lai, M. Xie, and D. N. P. Murthy, "A modified Weibull distribution," *IEEE Transactions on Reliability,* vol. 52, no. 1, pp. 33-37, March 2003, doi: 10.1109/tr.2002.805788.

[17] O. O. Aalen and H. K. Gjessing, "Understanding the shape of the hazard rate: A process point of view," *Statistical Science,* vol. 16, no. 1, pp. 1-14, Feb 2001.

[18] S. K. Maurya, A. Kaushik, S. K. Singh, and U. Singh, "A new class of distribution having decreasing, increasing, and bathtub-shaped failure rate," *Communications in Statistics-Theory and Methods,* vol. 46, no. 20, pp. 10359-10372, 2017, doi: 10.1080/03610926.2016.1235196.

[19] Q. H. Duan and J. R. Liu, "Modelling a Bathtub-Shaped Failure Rate by a Coxian Distribution," *IEEE Transactions on Reliability,* vol. 65, no. 2, pp. 878-885, Jun 2016, doi: 10.1109/tr.2015.2494374.

[20] W. J. Roesch, "Using a new bathtub curve to correlate quality and reliability," *Microelectronics Reliability,* vol. 52, no. 12, pp. 2864-2869, Dec 2012, doi: 10.1016/j.microrel.2012.08.022.

[21] J. Navarro and P. J. Hernandez, "How to obtain bathtub-shaped failure rate models from normal mixtures," *Probability in the Engineering and Informational Sciences,* vol. 18, no. 4, pp. 511-531, 2004.

[22] S. Rajarshi and M. B. Rajarshi, "Bathtub distributions - a review," *Communications in Statistics-Theory and Methods,* vol. 17, no. 8, pp. 2597-2621, 1988, doi: 10.1080/03610928808829761.

[23] M. V. Aarset, "How to identify an bathtub hazard rate," *IEEE Transactions on Reliability,* vol. 36, no. 1, pp. 106-108, Apr 1987, doi: 10.1109/tr.1987.5222310.

[24] K. L. Wong, "The bathtub does not hold water any more," *Quality and Reliability Engineering International,* vol. 4, no. 3, pp. 279-282, 1988, doi: 10.1002/qre.4680040311.

[25] H. T. Zeng, T. Lan, and Q. M. Chen, "Five and four-parameter lifetime distributions for bathtub-shaped failure rate using Perks mortality equation," *Reliability Engineering & System Safety,* vol. 152, pp. 307-315, August 2016, doi: 10.1016/j.ress.2016.03.014.

[26] D. N. P. Murthy and R. Jiang, "Parametric study of sectional models involving two Weibull distributions," *Reliability Engineering & System Safety,* vol. 56, no. 2, pp. 151-159, May 1997, doi: 10.1016/s0951-8320(96)00114-7.

[27] G. A. Klutke, P. C. Kiessler, and M. A. Wortman, "A critical look at the bathtub curve," *IEEE Transactions on Reliability,* vol. 52, no. 1, pp. 125-129, 2003, doi: 10.1109/TR.2002.804492.

Asynchronous Vehicle Control System Based on Integrated Driver Support Algorithm

Damian Petrecki
Department of Computer Science and Management
Wrocław University of Science and Technology
Wrocław, Poland
e-mail: damian.petrecki@pwr.edu.pl

***Abstract* — The paper describes a proposition of a driver support system composed of multiple independent processes producing discrete outputs and consuming continuous inputs, with a shared interpolate process. The research rejects multiple controllers handling different areas with the same actuators in favor of single but both multi-criteria and asynchronous decision-making system. This way, a decision-making problem has been limited, and a big data processing and control conflict hazard has been eliminated, keeping high vehicle performance, and lowering physical system complexity. This is an absolutely novel solution, very different than existing multi-domain in-vehicle controllers, due to new tasks division and processes synchronization approach. The results obtained during the simulation-based experiments show a very promising safety and comfortable ride.**

Keywords – integrated driver support system; multi-domain controller; continuous-time control; asynchronous algorithm, Integrated Driver Support Algorithm.

I. INTRODUCTION

A modern car can be controlled by a driver together with multiple control systems. Some of them just support the driver and some can take over entire control – at all times or in specific situations. It is getting increasingly critical for the automotive industry, customers and even governments and other lawmakers. It can be seen even in automotive marketing actions – new safety systems and drivers' assistance systems are boosted over new engines, better comfort, or practicality.

The drawback is that all these systems are created and implemented separately, often by different companies at different times, and simply not designed to work together. Therefore, a lot of devices are duplicated, multiple controllers with kilometers of cables are used, and control conflict hazard has to be maintained.

A proposed solution to this problem is to integrate all such systems into a single one, with one controller and a novel algorithm to rule all in-vehicle actuators.

The paper describes the algorithm – a new way to decompose the driver support problem, not into stability control, slipping wheels issue, extreme situation handling, etc., but into data acquisition, trajectory calculation, and control execution, which provides comparable results: a safe and comfortable ride. The algorithm has been called Integrated Driver Support Algorithm. The solution is literally a heuristic algorithm performing the vehicle control task, basing on driver's reference input and awareness of the surroundings, exclusively producing actuator signals for all actuators in the systems. This way, a vehicle equipped with the proposed solution can use a single, centralized computer system that eliminates multi-system interferences. Moreover, the vehicle can be easily maintained, including over-the-air software-based tuning, updating, and introducing new features without adding new physical sensors and controllers or modifying existing ones. What is essential, there is no ability to bypass the system by a driver, so it cannot be called a typical decision support system.

The paper is an extended version of [1], with a new, novel surrounding analysis method, and also supplemented by latest tests results, more descriptive explanations of previous ones and at last – a new, established algorithm name.

The paper structure is as follows. Section II shows a classic approach and currently popular research topics in the automotive area. The algorithm is described in Section III, which is followed by the presentation of the conducted simulation and its results in Section IV. The results of the simulations are evaluated in Section V. Comments on further work given in Section VI complete the paper.

II. STATE OF THE ART

It is hard to point-out a modern and adequately justified driver support system. There are well-known standard safety systems, like Anti-Lock Braking System (ABS) [2] or Electronic Stability Control (ESC) [3], but the mainstream is developed under non-public licenses or even as companies' secrets.

Nevertheless, we can observe modern vehicles' behavior and reach a conclusion how such systems work. Let us consider a case study, a widespread situation, well known from everyday driving – a driver wants to launch rapidly with front wheels turned, like when entering the flow of traffic. A modern car equipped with typical safety systems would involve a lot of these to influence the same parameter – wheels' speed. Engine Management System (EMS) [4] uses the engine to raise it, Acceleration Slip Regulation (ASR) [5] reduces it, active differential differentiates it, ESC applies brakes to avoid slipping, and ABS limits this

brake action. This description omits other safety systems also able to use brakes when triggered, e.g., some surrounding aware collision prevention systems.

There is also a common issue for modern, existing solutions, briefly mentioned in the introduction – a hardware duplication. Modern vehicles are equipped with multiple sensors that measure the same parameters or areas, like a camera for traffic signs recognition, another one for lane assist system [6], another for pedestrian avoidance system, and another intended to control headlights – all directed ahead of the vehicle. With each sensor, there is separate wiring and, of course, a processing unit. This way, the complexity of the system raises, along with the cost, mass, failure probability and maintenance difficulty.

On the other hand, still, the most popular, related scientific topics are vision and perception [7][8], traffic models [9][10], or accident preventions [11][12]. Such papers are made to improve existing systems with better performance, lower cost, or extra features.

As it has to be mentioned, there is another, significant direction in the current automotive-related research – autonomous driving [13-15]. The aim of autonomous vehicles is to replace the driver by a surrounding and traffic-aware, intelligent algorithm or algorithms. To achieve it, a lot of different safety features were introduced – lane support assists, active cruise control, GPS and map-assisted localization mechanisms, traffic sign recognition system, pedestrian tracking system, etc. A lot of new papers are being produced around this issue and its different dimensions, problems, and perspectives. According to statistic research [16], drivers cause the vast majority of all accidents, so this approach is justified and has a lot of advantages. It is worth to mention that the solution proposed in this paper can be a potential, very convenient base for an autonomous driving system, but is not designed to be the base of it from the beginning.

There is also a new approach to create multi-domain controllers to handle several issues together. We can safely assume that modern stability control systems, even the companies-secret ones, are developed as single controllers that can control an engine and each brake separately, and realize ESC, ABS, ASR and another similar systems' functions within the same decision-making processes.

The next step is to integrate in-vehicle IT infrastructure and to create a single system to rule all actuators in a vehicle with knowledge coming from all sensor. One of the solutions which is under development now, is AUTOSAR [17]. It is a hardware-software solution intended to compose

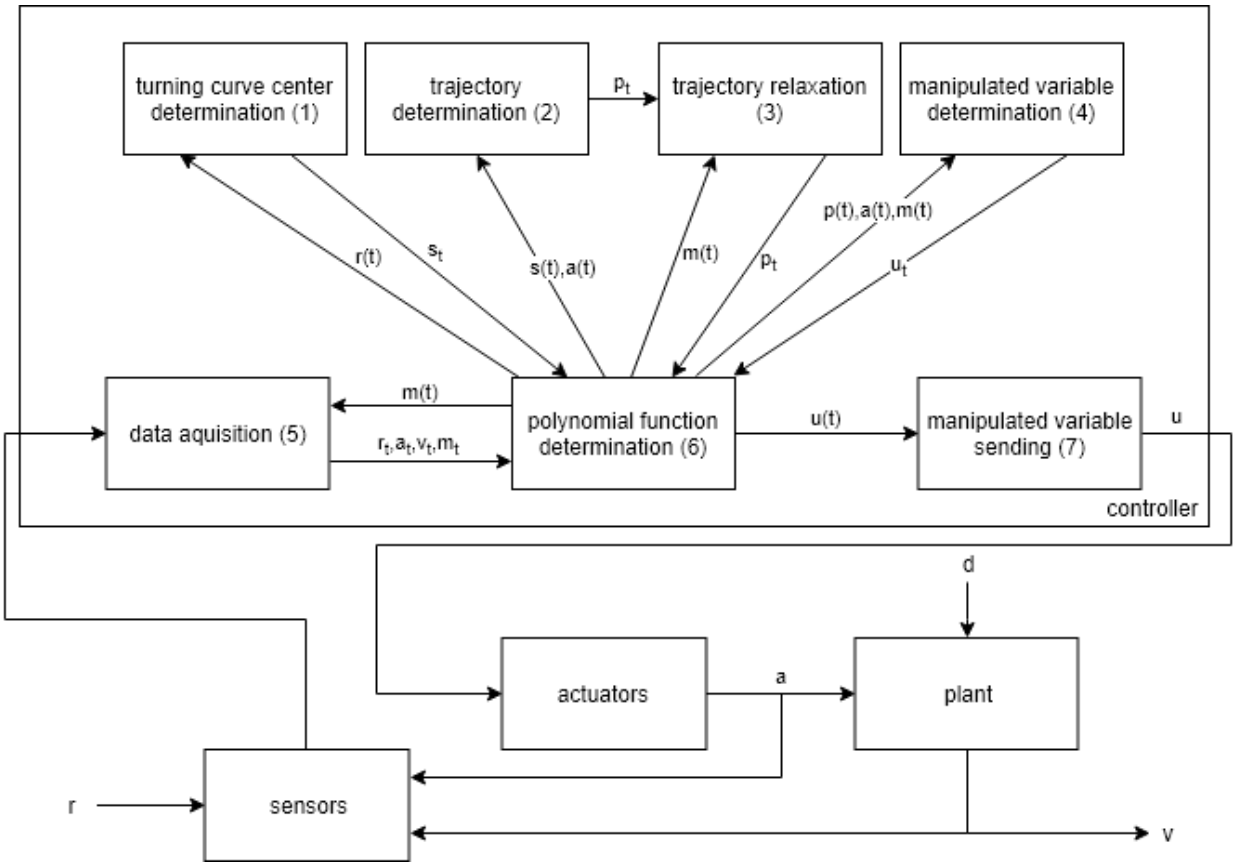


Figure 1. Control system schema

TABLE I. SYMBOLS USED IN FIGURE 1

symbol	description
r	reference input form an user
a	actuators state
d	external distortions
v	behavior measured via sensors (cameras, radars, accelerometers, etc.)
m	surrounding map – list of measured objects with its position and classification and parametrization results
u	control vector
s	center of turn
p	vehicle position p_t/trajectory *p(t)*

multiple control systems into a single one. This project notable expands the scope of a currently used, popular in-vehicle computer network, called CAN [18] or FlexRay [19], which is the only integration point now.

III. ALGORITHM GENERAL DESCRIPTION

This research presents a different perspective to driver support systems – it is an integrated system, and the main goal is not developing a better perception system, a more precise model, or a smarter autonomous driver-replacement, but presenting a new way to compose different, existing solutions to achieve high performance (in a way of vehicle safety and comfort) while lowering the computing power and system physical complexity at the same time.

The idea of algorithm is shown in Figure 1 with symbols explained in Table 1. When modeling using the black-box method and ignoring the controller's structure, the presented solution seems to be very similar to typical control systems. It reads reference input *r* from the user (mainly steering wheel and pedals positions), the vehicle behavior in its surrounding (using cameras and radars) and vehicle-related data (using accelerometers, thermometers, position sensors, etc.) *v* and also actuators state readers *a* and produces a control vector *u* consisting of all actuators manipulated variables: engine, linkage system, suspension, driveline, brakes. The only difference is the lack of time connection between inputs and the output.

When modelling using the white-box method, it can be seen that the algorithm refuses to use a popular algorithm chain architecture, based on iteratively producing new output data by calculating new input data with existing state data.

The algorithm consists of several processes instead. Each of them can be scheduled (triggered by time) or started by data incoming from a sensor. The outputs of all processes are discrete values (in one or more dimensions) shown as sequence elements with bottom index *t*, e.g., s_t. Input data (both starting processes and read during them) come from continuous-time functions stored in analytical, polynomial forms, shown as functions with *t*-argument, e.g., *s(t)*. It means a single, distinguished process (referred in Figure 1 as (6)) is introduced to build continuous-time functions from discrete sequences, which allows data interpolation and extrapolation. This way all data can be read between real measurements or calculations even after last ones without loosing accuracy. The process uses polynomial curve fitter method [20], accepts discrete values and timestamps, and produces a vector of polynomial coefficients. This way, a very specific storage is introduced that stores discrete variables and provides analytical functions as its output.

A single-dimensional data acquisition process is proposed (5). It reads and stores input values from input devices r_t and in-vehicle sensors, it reads actuators' states a_t, like suspension status, accelerations, engine status, etc., and simple (non-matrix) measurements from v_t. Each variable is handled by a separated thread.

The second part of process (5) handles the surrounding data v_t and is the most complex one. This is a complex part of the data acquisition process. This is the only case when the matrix data (distances of radars or bitmaps from cameras) must be handled. The process is triggered for each input from each signal separately. The result is a 3D model of the vehicle surrounding consisting of a set of classified objects, in the form of objects' shapes (3D line segments), class and positions, so data size is significantly decreased. Due to long processing time, the output of this process is stored with the input data appearance timestamps.

The surrounding analysis process is the biggest challenge related to the research. It needs a separate algorithm that accepts various formats of input data arriving at unpredictable time (cameras, radars) with, optionally, the already known 3D surrounding model *v(t)* to update the model as its output. The Long Short-Term Memory (LSTM) [21] neural networks were considered as the most promising way to solve this problem, but it turned out that a very typical neural network used in some unusual form is a better solution.

Therefore, TensorFlow [22] platform is used to face this issue, but instead of using neural networks to classify objects, they are used to localize objects in the vehicle surrounding. Each network is trained to return a proper position of an object of a single class (like a specific model of vehicle, or a tree of a known shape) or zero when no object of this class is found. It means that each known class requires a separately trained network which results in a very long training process. The process was automated using scripts in the simulation environment during the research.

When the algorithm is being started, all known networks have to be run to analyze the vehicle surrounding, but after that, the existing surrounding model is used to limit the number of networks, using a list of already found or expected objects.

The process of the next type calculates the desired trajectory using the surrounding knowledge, vehicle-geometry model, and input data. It is split into several sub-processes, without any time-synchronization:

- The first sub-process referred to (1) calculates the center of the turning curve (if any) in the vehicle-centered coordinate system, using speed, steering wheel position, and vehicle geometry.
- Sub-process (2) calculates the desired vehicle positions p_t in the future, which means the desired vehicle trajectory.
- Sub-process (3) uses a genetic relaxation algorithm [23] and the surrounding knowledge v_t to improve the trajectory to avoid accidents, lowering external objects hit possibility. Please note this process can change the trajectory in any way, e.g., by decreasing speed or changing the turn, and its behavior is unpredictable. This is the only process that reads its input directly from the other process, not the storage.

The next process (4) uses the trajectory *p(t)* to calculate control values for all executors u_t. For example, it calculates each wheel speed and turn, followed by the determination of engine power, braking force, linkage system, and differential parameters. This process bases on a mathematical model of the vehicle. It cannot be assumed that the model is utterly reliable and precise, but as it was realized during experiments, inaccuracy does not affect the evaluation of the algorithm. The model ignores the vehicle's body stiffness, uses Pacejka "magic formula" tire model and simple stiffness/dumping suspension model and it is still good enough for purposes of the algorithm. Moreover, using a more complex model, better reflecting a real vehicle behavior, impacts negatively on computation power usage but does not significantly improve overall algorithm evaluation results.

Calculated manipulated variables are being sent to the vehicle by own sender processes (7) (one process per variable), which read data from storage, not from the processes that actually generated them. Please note when using such algorithm architecture, there is no guarantee that output data is calculated using lastly collected input data. The delay can be relatively big (up to over a dozen measurements), and there is no mathematical proof that it does not affect the overall results. However, during the experiments, such a negative impact was not observed even during the worst-case scenarios, with rapid, unpredictable condition changes.

Please also note that processes (2) and (3) are shown in Figure 1 as two distinct parts, because each of them has its own interface and can be replaced by any other algorithm that fulfills it. On the other hand, it is a single process that uses the storage at its input and output and produces the trajectory *p(t)* in two steps. In the same way, process (5) is shown as a single one because it has a single interface, but it effectively consists of two internal sub-processes to handle different formats of incoming data.

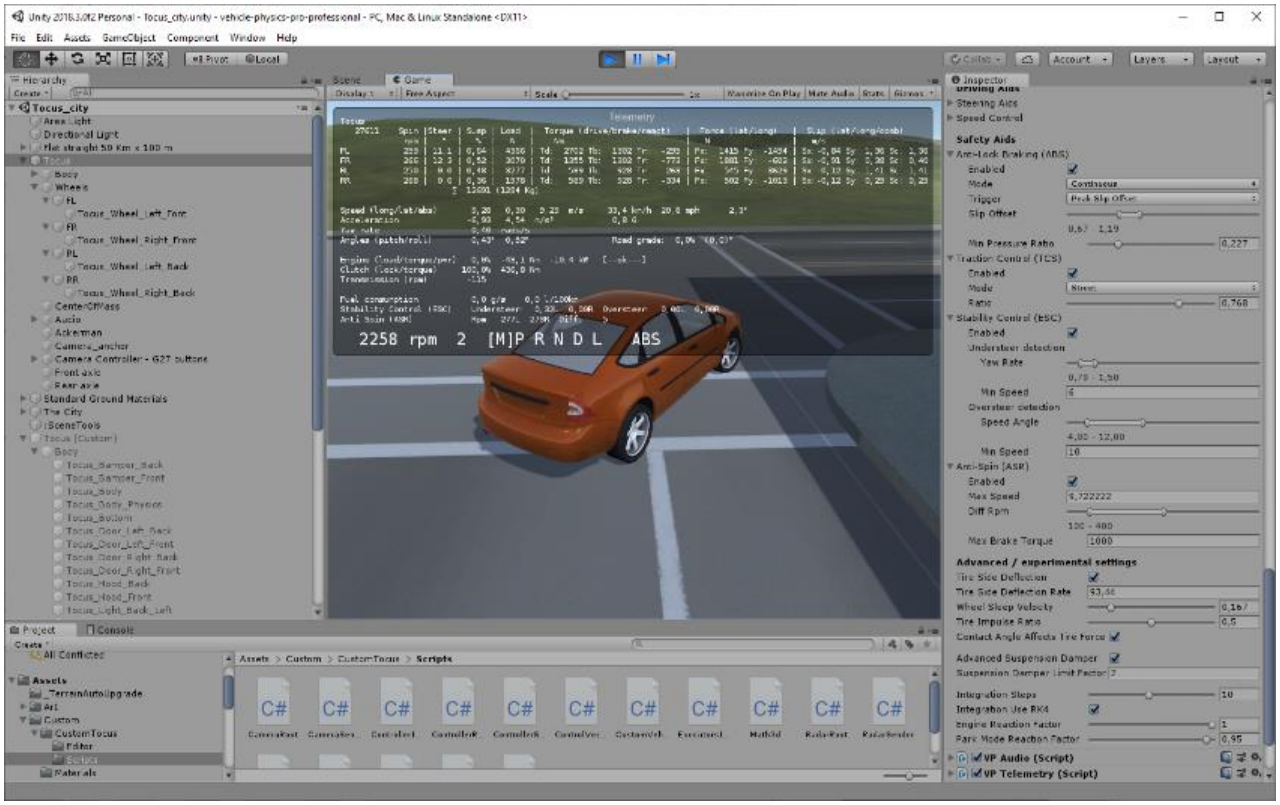


Figure 2. Simulator view (90-degree test)

TABLE II. MOOSE TEST RESULTS

enter speed	reference solution			proposed solution		
	A	*B*	*C*	*A*	*B*	*C*
80	1.6	-4	170	0.1	2	91
80	1.8	-7	172	0.1	-1	84
80	1.7	-8	168	0.1	0	83
80	1.6	-5	192	0.2	1	86
100	1.6	-6	99	0.1	2	100
100	1.6	-6	97	0.2	11	87
100	1.7	-10	78	0.1	-2	81
100	1.7	-6	89	0.1	0	82
120	1.8	-9	145	0.2	1	92
120	1.5	-9	150	0.2	4	91
120	1.7	-2	154	0.2	4	97
120	2	-4	140	0.2	-1	95
140	failed			0.1	2	110
140	1.6	-40	180	0.4	2	101
140	2.1	-36	165	0.2	4	105
140	1.9	-38	79	0.3	1	87
160	1.9	-48	145	0.2	-2	89
160	1.8	-60	138	0.3	1	90
160	2.2	-40	139	0.3	2	98
160	failed			0.3	2	115
180	failed			0.4	2	97
180	2.1	-68	165	0.3	2	95
180	failed			0.3	1	96
180	failed			0.2	1	97
200	2	-20	66	0.2	2	94
200	Failed			0.3	1	119
200	1.9	-30	81	0.4	-6	126
200	failed			0.3	-1	83

TABLE III. 90-DEGREE TURN TEST RESULTS

enter speed	reference solution			proposed solution		
	A	*B*	*C*	*A*	*B*	*C*
10	0.2	2	253	0.2	-1	76
10	0.1	1	268	0.2	1	99
10	0.4	0	342	0.1	2	108
10	0.4	-2	268	0.2	0	104
20	0.8	1	372	0.2	-1	76
20	0.7	2	371	0.2	2	108
20	0.9	0	365	0.3	1	104
20	1	-1	312	0.2	0	98
30	1.2	2	290	0.2	2	96
30	1.2	1	246	0.2	1	94
30	1.1	-1	256	0.2	2	88
30	1.3	3	267	0.1	3	98
40	1.5	-10	160	0.2	-3	198
40	1.6	-8	381	0.2	-4	197
40	1.5	-12	271	0.3	-4	178
40	failed			0.4	3	174
50	1.9	-28	450	0.3	3	149
50	failed			0.2	-24	324
50	failed			0.3	-23	354
50	failed			failed		
60	failed			0.4	-38	450
60	failed			0.4	-39	450
60	failed			failed		
60	failed			0.4	-42	450

The most important idea behind the algorithm is the absolute lack of time synchronization between its input and output. Each process is being run separately and uses data extrapolated or interpolated from other processes, no matter the age of the source values or the last polynomial calculation time.

The second most important idea is to allow replacing processes with similar ones, that use the same interfaces. For example, this way the interpolation algorithm can be replaced by trigonometric one or TensorFlow can be replaced by some computer vision-based algorithm, without affecting the algorithm architecture, even with simple software upgrade of a vehicle.

The third main idea is to not focus on well-known automotive-related issues, like preventing wheel slipping or speedway lane recognition, but to use computer science knowledge, a decision making algorithm and a robotic-like approach to monitor and improve a driver expertise and intuition.

It must be noted that all functions calculated by the algorithm can be used by external processes, not related to vehicle control, like headlight control, climate control, comfort features, etc., but it is not part of this research.

IV. CURRENT RESULTS

The presented solution has been tested in different scenarios, and the current results are presented.

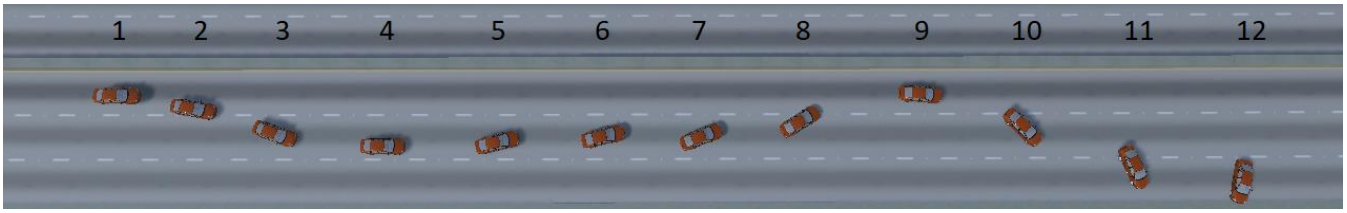


Figure 3. Moose test visualization for the reference vehicle and enter speed 140km/h (test failed)

A. Simulation environment

All experiments are conducted in Unity3D [24] environment with Vehicle Physics Pro (VPP) [25]. Unity3D is responsible for communication with an operating system driver of Logitech G29 steering wheel [26], rendering visual interpretation of simulated rides (shown in Figure 2 and Figure 3), and fundamental Newton physics. VPP is responsible for the vehicle simulation, including dependencies between in-vehicle physical subsystems, tires and suspension behavior, and standard active safety systems. VPP is also a source of the reference vehicle used, with all safety systems already included and configured. VPP is, in general, a pre-compiled library, so a lot of its mechanisms are unknown. Its realism is not verified, just considered to be sufficient to compare two vehicles in the same conditions. The validity of results obtained on the basis of a simplified environment for a corresponding real-world environment is the principal assumption of the research. The experiments' results are read from the telemetry panel provided by VPP (Figure 2) and from controller application. All data are stored during the tests in text files and analyzed offline.

B. Test and reference vehicles

The reference vehicle is built using VPP components only. It has an active suspension, 4-wheel steering, automatic gearbox, 4-wheel drive with active differential, and following active safety systems: ABS, Traction Control System, ESC, ASR. Most of its implementation is hidden and unknown but is calibrated using built-in configuration

TABLE IV. MOOSE WITH OBSTACLE TEST RESULTS

enter speed	reference solution			proposed solution		
	A	*B*	*C*	*A*	*B*	*C*
120	1.4	-4	145	0.2	0	87
120	1.5	-5	155	0.2	0	87
120	1.7	-6	149	0.3	4	89
120	1.9	-8	141	0.2	-2	89
140	1.6	-29	180	0.3	-1	98
140	2.0	-32	180	0.2	-3	99
140	failed			0.1	4	100
140	1.9	-32	81	0.2	5	89
160	2.0	-65	145	0.2	-1	110
160	1.9	-54	154	0.2	8	98
160	1.9	-53	163	0.3	4	95
160	failed			0.4	-4	110
180	failed			0.4	-5	101
180	2.2	-66	153	0.3	2	104
180	failed			0.3	-5	97
180	failed			0.3	-5	109
200	failed			0.2	-6	94
200	failed			0.5	6	111
200	failed			0.4	-6	132
200	failed			0.4	3	87

TABLE V. CRASH TEST RESULTS

enter speed	reference solution	proposed solution
60	avoided	avoided
60	avoided	avoided
60	avoided	avoided
60	avoided	avoided
80	wall	avoided
80	avoided	avoided
80	avoided	avoided
80	wall	avoided
100	avoided	avoided
100	wall	avoided
100	wall	avoided
100	following	avoided
120	wall + following	wall + following
120	wall + following	avoided
120	wall + following	avoided
120	following	avoided
140	other	other
140	other	other + following
140	other + following	other + following
160	other + following + wall	other
160	other + wall + following	other + following
160	wall + following	other
160	wall + following	other + wall

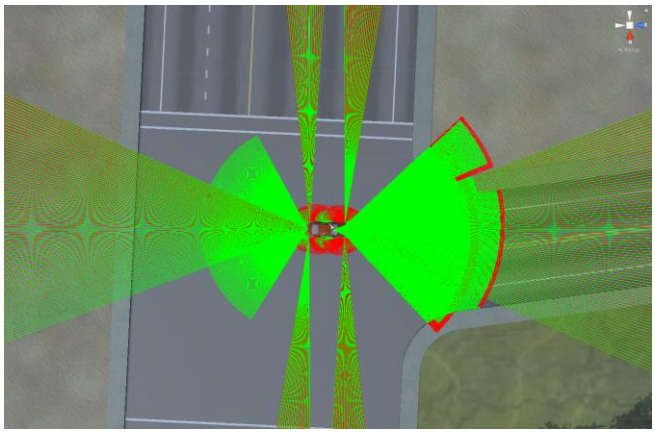


Figure 4. Lidar coverage (red lines mean no-hit rays)

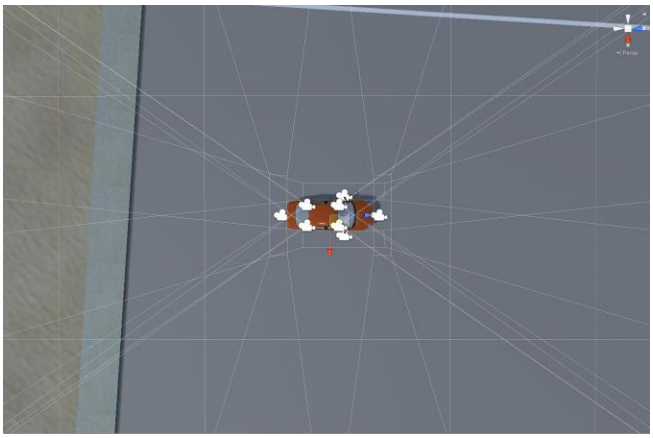


Figure 5. Camera coverage

panels, visible on the right part of Figure 2. Some systems, like active differential, were not available with VPP so have been implemented manually for this research.

The test vehicle has the same 3D model, physical parameters (weight 1200kg and equal weight distribution per wheel, wheels localizations, engine power/torque curves, etc.) and abilities (4-wheel drive, 4-wheel steering, controllable transmission, differential, and suspension). So, it is the same vehicle equipped with a different control system.

The difference is that in the test vehicle, the input from a driver is not sent to the vehicle itself but transferred to an external application implementing the proposed algorithm. All in-vehicle and surrounding-related sensors data are handled in the same way. The application sends back control variables for each actuator separately in separate threads. In this vehicle, there is no other driver support system implemented.

C. Performance results

All experiments are conducted using i7-7700k 4.2Ghz processor, 16GB RAM, SSD hard drive, and Windows 10 64bit operating system. Both simulation (Unity3D) and control (external application) are performed on the same machine because its performance is sufficient for current test scenarios. RAM usage never exceeds 10GB, and CPU load is always below 20% when simulation framerate 40fps is preserved.

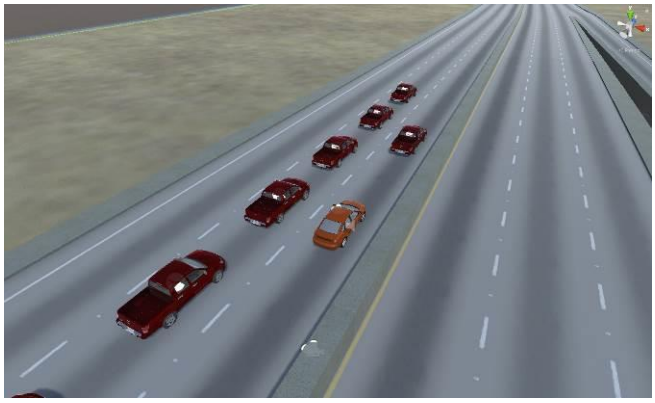


Figure 6. Experiment 4 scenario

The situation changes when all external sensors (21 LIDAR's and 8 cameras – see Figure 4 and Figure 5) are running. Then the simulation occupies about 4GB more RAM (which is still irrelevant), but exhausts all CPU abilities, reducing simulation framerate to 10-15fps (depending on a scenario). There are two ways to face this issue – to split simulation execution and control calculation to different computers using standard TCP/IP and UDP/IP network stacks or to disable several sensors simply. For the purpose of this paper, the second option is chosen, and all sensors directed backward are disabled.

D. Dynamic experiments

Five different experiments have been performed. All test rides have been conducted 4 times, each with the same driver and the same conditions. In Tests 1-3, vehicle roll angle A (degree), speed change B (km/h), and maximum steering wheel angle C (degree) have been evaluated. Lower roll angle means better comfort. Lower loss of speed means higher safety (shorter maneuver time) and also better efficiency (energy loss). Lower steering wheel rotation angle is considered as sportier and even safer behavior, allowing the driver to turn faster when holding the steering wheel with both hands all the time. During Test 4, both vehicles should avoid collisions and stay on the road, so the evaluation is descriptive. Possible values are: "avoided" (no crash), "wall" (crashed into a wall), "following" (crashed into the following vehicle), "other" (crashed into a vehicle on the other line) or a combination of them. In the last test, the most important evaluation parameter is minimum D and maximum E tires slip (m/s). Lower slip means better handling and safer ride. The test is passed when car fits the 4.3m lanes during the entire test.

1) Moose test

The test scenario requires a rapid change a lane and return to the original one on a straight road with velocity in range 80-200km/h (changing by 20km/h).

Results of the test are shown in Table II and Figure 3.

TABLE VI. LONG TURN TEST RESULTS

enter speed	reference solution		proposed solution	
	D	***E***	***D***	***E***
20	0.01	0.23	0.07	0.07
20	0.02	0.25	0.06	0.07
20	0.01	0.22	0.05	0.07
20	0.01	0.23	0.06	0.06
30	0.06	0.4	0.1	0.11
30	0.05	0.42	0.09	0.11
30	0.06	0.41	0.09	0.12
30	0.07	0.4	0.11	0.13
40	0.18	0.72	0.27	0.31
40	0.2	0.7	0.22	0.25
40	0.19	0.7	0.25	0.3
40	0.17	0.71	0.27	0.3
50	0.22	1.81	0.2	0.21
50	0.21	1.9	0.23	0.26
50	0.24	1.86	0.25	0.29
50	0.24	1.78	0.21	0.24
60	0.25	2.59	0.26	0.37
60	0.25	2.72	0.25	0.28
60	0.27	2.58	0.25	0.27
60	0.26	2.58	0.26	0,29
70	0.26	2.42	0.21	0.25
70	0.25	2.44	0.22	0.25
70	0.26	2.53	0.21	0.26
70	0.26	2.51	0.25	0.28

2) 90-degree turn

The test scenario assumes the turn right on a 90-degree intersection with velocity in range 10-60km/h (changing by 10km/h).

Results are shown in Table III.

3) Moose test with an obstacle

During this test, the vehicle should avoid a collision. The obstacle of a known class appears from the right side (like a vehicle coming in from a side road) when riding with velocity in range 120-200km/h (changing by 20km/h), on a two-lane road. No incoming traffic is taken into consideration. The aim of the test is to avoid a collision and return to the lane. The test is passed when a collision is avoided, and the vehicle fits in two 4.3m lanes during the entire test ride.

Experiment results are shown in Table IV.

4) Crash test

During the test, a vehicle rides behind another one on the right lane of a two-lane road. The left and right lane are crowded and walled, respectively (Figure 6). All vehicles start each test located in the middle of the proper lanes. The vehicle in front of the test one stops rapidly (with infinitive decelerating force and traction). At this moment, both vehicles (the test one and the one in front of it) are separated from each other by a distance value 2 times smaller than velocity, with unit conversion from velocity (km/h) to distance (m). It means 25m gap for 50km/h, 50m gap for 100km/h, etc. Tested velocity is in range 60-160km/h (changing by 20km/h).

The aim of the test is to avoid a collision or reduce an impact when possible. Results are given in Table V.

5) Long turn

Now a vehicle rides around a circle with a constant radius of 20m and speed in range 10-70km/h (changing by 10km/h). In this test, the capability of separately controlling all-wheel speed is shown.

Experiment results are shown in Table VI and Figure 7.

V. CONCLUSIONS

When analyzing results, some considerations arise. The reference vehicle has failed in 29% and the tested one in 4% of all 1st and 2nd tests' trials. This is the main proof of improved safety in the presented solution. Secondly, the maximum roll of the test vehicle is limited to less than 0.5 degrees no matter of conditions, for all trials. The reason is that the anti-roll bars work pro-actively, reacting to turn and speed, not to the roll itself. This behavior proves that the comfort of the ride improved.

The next thing to notice is that the maximum steering wheel rotation angle in the test vehicle is significantly lower and fits into 90 degrees for most cases. The cause is the steering wheel ratio being adjustable in an extensive range, due to the lack of physical connection (even simulated one). The function converting wheel angle and speed to the position of the center of a turn is adjusted to lower the minimum turn radius at high speed when rapid turns are impossible anyway due to vehicle momentum.

The next observation is that the presented vehicle does not slow down during most of the tests, except the ones, when preserving speed is impossible, due to high vehicle inertia. The cause is that the driver does not press the brakes, so all trajectories are calculated for the same speed. Stability is preserved with an active differential that transfers proper speed to all wheels to avoid a slip, with the stiff connection between wheels and engine and without using brakes. This way, a maneuver can be finished faster, and the engine is never stalled by brakes, which also improves safety.

On the other hand, the reference vehicle uses brakes to preserve comparable stability which causes a significant loss of speed.

As it is seen in Figure 3, vehicle position 6, dangerous loss of control over a vehicle can happen even when the vehicle is equipped with ABS and ESC. No such thing has happened during all experiments for the test vehicle, for any

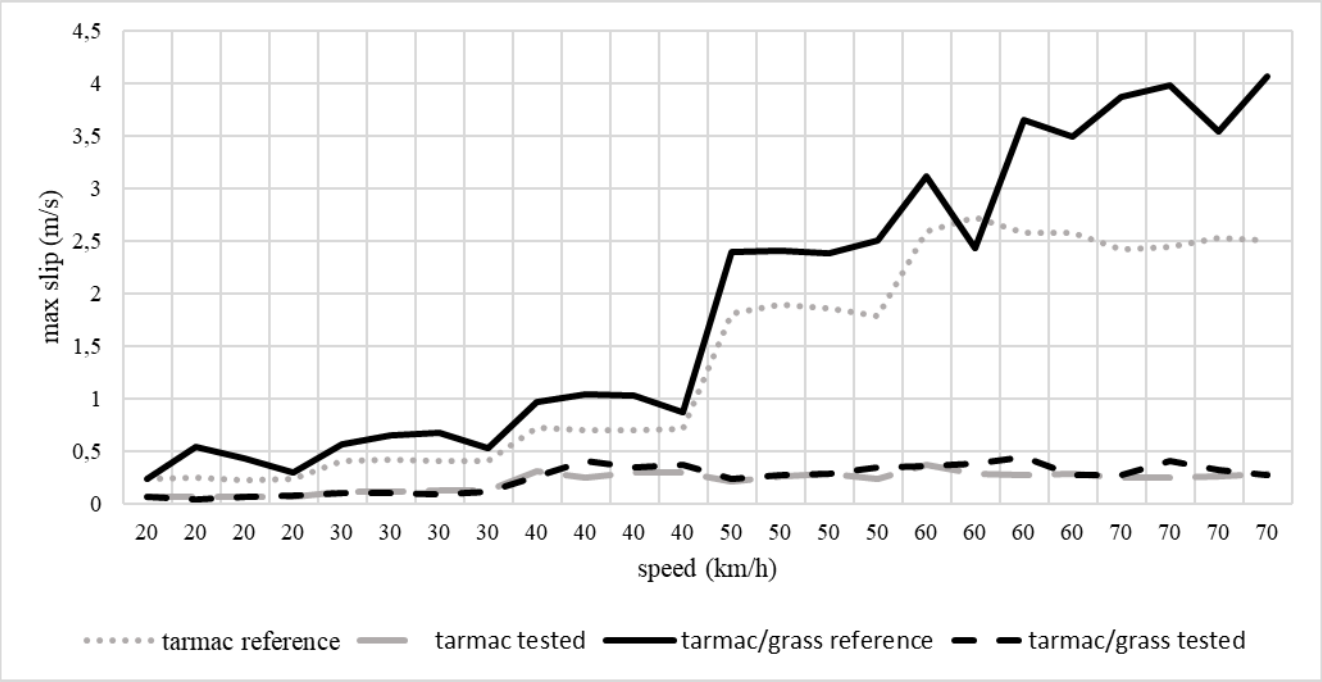


Figure 7. Slip test on different surfaces

scenario. It proves that without awareness of the surroundings, the reference vehicle follows driver's commands, even when the driver leads the vehicle to an accident.

Test 3 coincides, in general, with Test 1. It proves that awareness of the surroundings does not affect safety in a negative way. Neither hazardous nor dangerous situations were observed during the test vehicle's rides. What is important, the driver did not see any trajectory changes, even when it occurred. It proves seamless but safe trajectory changes calculated in the relaxation process.

Test 4 has delivered very interesting results. For all tests where a crash could be avoided, the test vehicle managed to avoid it. The driver reported that the car triggered braking before hitting the pedal or even without it. This behavior proves that awareness of the surroundings has been used to calculate and execute an accident-free trajectory. The action of the test vehicle after a first hit is the next interesting observation. In such situations, the entire algorithm is still working and trying to avoid another accident. The test has passed for the majority of repetitions. The mathematical model of the vehicle, that is not working well when vehicle inertia is disturbed after a hit, is a probable reason for the noticed failures of the test. A damage of a vehicle has not been comprised by the model. Almost all crashes up to 120km/h have been avoided by the test vehicle that is another advantage of the proposed algorithm. The vehicle tried to fit among the foregoing car and cars located on the second lane. The only failed result occurred when a driver has intentionally turned right during the trial.

On the other hand, the behavior of the reference vehicle has strictly depended on driver's actions during all rides. It has to be mentioned that the driver had also tried not to hit obstacles directly but to fit between them and in some tests (no. 16 and 18) had touched obstacles very gently even though high velocity. In similar situations, the test vehicle had crashed very hard each time when a collision was inevitable. This behavior is caused by a fallback mechanism implemented in the algorithm. When a collision probability is certain (trajectory relaxation cannot find any accident-free trajectory within a configured timeout), the unchanged path is returned from the process (3). It means that the vehicle follows the driver's trajectory without any changes and the driver may not be prepared for that.

In the last test, the minimum slip, occurring for inner wheels, is comparable for both vehicles, but active differential implementation offered by VPP is not as effective as the tested one. Moreover, in different conditions, when outer wheels travel on grass instead of tarmac, the differences between the reference and test vehicles are even bigger (Figure 3). The reason is, again, that tested standard safety systems react by breaking wheels already slipping, and the test one controls the behavior of the wheels proactively, calculating its speed before any slip occurs using mainly engine and differential, not brakes. This result also proves higher ride safety.

VI. FUTURE WORK

Although the current results are promising, a lot of work is planned. For now, all processes are triggered by data or time. The event-based trigger (rapid condition change) is planned. Besides that, disruption analysis with fuzzy functions [27] usage will be introduced, to replace all

continuous-time functions with fuzzy equivalents to improve the overall evaluation. The full assessment will be conducted with more test cases and more drivers to find more edge-case scenarios to improve. And lastly, all processes implementations of the method use very simple algorithms so far, but they are designed to be replaceable, so the best combination is to be found. The first element to replace is the trajectory relaxation's objective function, which should minimalize possible accident impact instead of just avoiding it.

To make the evaluation of the algorithm fairer and more unprejudiced, a surrounding aware, automatic braking system has to be introduced to the reference car.

Future experiments will be conducted using two computers with a direct network link.

Physical experiments with real vehicles are not planned so far. Very sophisticated, highly equipped vehicle (with active differential, suspension, etc.) with an open-access available to all in-vehicle actuators and sensors and also a large set of extra sensors are needed, which makes such experiments too expensive for the current stage of research. This kind of research is possible after the full simulation evaluation.

REFERENCES

[1] D. Petrecki, "Asynchronous Vehicle Control System Basing on Analytical Continuous-Time Functions", in: The Eighth International Conference on Advances in Vehicular Systems, Technologies and Applications (VEHICULAR 2019), Rome, Italy, June-July 2019.

[2] J. Ernst, "Mercedes-Benz and the invention of the anti-lock braking system: ABS, ready for production in 1978", in: Daimler Communications, July 2008.

[3] A. van Zanten, "Bosch ESP Systems: 5 Years of Experience" in: SAE Technical Paper 2000-01-1633, January 2000.

[4] T. Denton, "Automobile Electrical and Electronic Systems", Chapter 10, Routledge, July 2007.

[5] D. Hoffman, "The Corvette Acceleration Slip Regulation (ASR) Application with Preloaded Limited Slip Differential," SAE Technical Paper 920642, February 1992.

[6] E. D. Dickmanns, "Dynamic Vision for Perception and Control of Motion", chapter 7, Springer, June 2007.

[7] M. Knorr, "Self-Calibration of Multi-Camera Systems for Vehicle Surround Sensing", KIT Scientific Publishing, December 2018.

[8] B. Ranft and C. Stiller, "The role of machine vision for intelligent vehicles", in: IEEE Transactions on Intellifent Vehicles, Vol. 1, pp. 8-19, March 2016.

[9] M. Z. Liu and M. Sun, "Application of Multidimensional Data Model in the Traffic Accident Data Warehouse", in: Applied Mechanics and Materials, Volumes 548-549, pp. 1857-1861, April 2014.

[10] D. Badura, "Prediction of Urban Traffic Flow Based on Generative Neural Network Model", in: Management Perspective for Transport Telematics. TST 2018. Communications in Computer and Information Science, volume 897, pp. 3-17, September 2018.

[11] Z. H. Ren, K. Zhang, L. X. Xue, and Y. L. Gao, "Mandatory Lane Change Model and Time Delay under Traffic Emergency Incidents", in: Applied Mechanics and Materials, Vols. 644-650, pp. 2627-2631, September 2014.

[12] H. Abut, J. Hansen, G. Schmidt, K. Takeda, and H. Ko, "Vehicle Systems and Driver Modelling", De Gruyter, September 2017.

[13] T. Nitsch, "Sensor Systems and Communication Technologies in Autonomous Driving", GRIN Verlag, April 2017.

[14] S. Liu, L. Li, J. Tang, S. Wu, and J. Gaudiot, "Creating Autonomous Vehicle Systems", Morgan & Claypool Publishers, October 2017.

[15] D. Vaishnavi, E. Sundari, T.V. Sangeetha, S. Shrinidhi, and P. Saravanan, "Design and Development of Computational Intelligence for Enhanced Adaptive Cruise Control Using Arduino", in: Applied Mechanics and Materials, Vol. 852, pp. 782-787, September 2016.

[16] N. Djurkovic, "Car Accident Statistics in The U.S. – 2020 Update", Carsurance, January 28, 2020, Accessed on: February 4, 2020. [Online] Available: https://carsurance.net/blog/car-accident-statistics/

[17] T. Scharnhorst and G. Reichart, "Progress on the AUTOSAR Adaptive Platform for Intelligent Vehicles", in: The Eighth International Conference on Advances in Vehicular Systems, Technologies and Applications (VEHICULAR 2019), Rome, Italy, June-July 2019.

[18] L. Renjun, L. Chu, L. Feng. "A design for automotive CAN bus monitoring system ". in: 2008 IEEE Vehicle Power and Propulsion Conference, Harbin, September 3-5 2008, pp. 1-5.

[19] R. Makowitz and C. Temple, "FlexRay-a communication network for automotive control systems", in: 2006 IEEE International Workshop on Factory Communication Systems, IEEE, 2006. pp. 207-212

[20] P. G. Guest, "Numerical Methods of Curve Fitting", Cambridge University Press, December 2012.

[21] A. Yenter, "A Multi-kernel Convolutional Neural Network with LSTM for Sentimental Analysis", ProQuest, 2017.

[22] https://www.tensorflow.org/, last retrieved, May 2020

[23] C. Cheng, C. Liu, and C. Liu, "Unit Commitment by Lagrangian Relaxation and Genetic Algorithms", in: IEEE Transactions on Power Systems, Vol. 15, pp. 707-714, May 2000.

[24] https://unity3d.com/, last retrieved May 2020.

[25] https://vehiclephysics.com/, last retrieved May 2020.

[26] https://www.logitechg.com/pl-pl/products/driving/drivingforce-racing-wheel.html, last retrieved May 2020.

[27] J. J. Buckley and E. Eslami, "Fuzzy Functions" in: An Introduction to Fuzzy Logic and Fuzzy Sets. Advances in Soft Computing, Physica, 2007.

A Business Model Analysis for Vehicle Generated Data as a Marketable Product or Service in the Automotive Industry

Frank Bodendorf and Joerg Franke
Friedrich-Alexander-University of Erlangen-Nuremberg (FAU)
Institute for Factory Automation and Production Systems
Erlangen, Germany
e-mail: frank.bodendorf@faps.fau.de
e-mail: joerg.franke@faps.fau.de

***Abstract*— Data-driven business models play a significant role in the digital transformation of traditional value-added industries. More and more existing and potential partners of automobile manufacturers show interest in the data generated by the vehicles. However, there is still no monetary value assessment to support decisions regarding the release of data. Traditional pricing approaches for material goods are based on cost, margin, and volume. However, these bottom-up calculation concepts are not applicable to digital goods. The background to this is, among other things, the uncertainty about the potential sales volume, the difficulty of cost splitting, and the high unit cost degression of digital goods. This paper provides a decision support for selling data to third parties as an intangible product. It introduces a concept that allows to value data generated by a motor vehicle in order to determine potential prospects and prices for sale. The evaluation model developed can be used to strengthen the car manufacturer's negotiating position towards potential data buyers.**

Keywords - Automotive Industry; Car Data; Business Model; Value Estimation

I. INTRODUCTION

The use of valuable data will fundamentally change competition in the future [31]. "The expected growth of the value pool from car data and shared mobility could add up to more than USD 1.5 trillion by 2030" [1]. Volume and quality of this "data treasure" will create strategic as well as operational competitive advantage [21].

Today, data is generated in large quantities by the vehicle, recording thousands of attributes. On the one hand, the vehicle user (driver) has the opportunity to enter data in on-board systems and "exchange" them for services. He/she is offered individually adapted functions, such as voice control, comfort settings when entering the car, navigational instructions in real time or other services [4][5]. On the other hand, a variety of sensors and computers in the vehicle, unnoticed by the driver, generates a steady stream of data, which among others serves for control purposes [14]. Examples are the anti-lock braking system or the automatic windscreen wiper and light regulation.

The data usage can be divided into nine purpose oriented categories [27]:

- Meeting regulatory and legal requirements (e.g., liability for material defects)
- Supporting marketing and advertisement (e.g., customer profiling)
- Assessing IT security (e.g., logging and monitoring)
- Improving technical processes (e.g., diagnostics and programming)
- Fulfilling terms of contract (e.g., new digital services and solutions)
- Innovating and developing products (e.g., monitoring and analytics)
- Ensuring road safety (e.g., traffic management)
- Transferring to third parties (e.g., car sharing)
- Facilitating vehicle use (e.g., autonomous driving)

All these categories have in common that value is created through the use of collected vehicle data. On the one hand, this value is reflected in technical or qualitative improvements as well as in cost reductions of the company's internal processes. On the other hand, the use of vehicle data can also lead to an economic improvement of the business results and in particular to an increase in turnover [32]. This may be a result of higher sales figures of products, i.e., manufactured vehicles, which are more attractive through data-based functions ("data infused products"). In addition, it is possible to offer certain data for sale as an end product itself [20].

This paper focuses on selling data to third parties. The demand for vehicle generated data depends on the benefit seen or expected by the buyer. From the perspective of the data provider, it is important to determine the value of the data in order to estimate the demand potential on an external market and to create appropriate pricing models [1].

II. EXISTING VALUATION APPROACHES

The generic value of a data product in sales situations cannot be determined by a benefit that has already been realized, since the data is not yet being used by the buyer at the time of the transfer. Therefore, an evaluation must be based on probable and potential benefits [20]. This value can be estimated by using qualitative and quantitative methodological approaches. For a corresponding systematic value determination of vehicle data, a number of existing evaluation methods are presented. This is initially done in a

tabular overview in Section II.A, followed by a more detailed description in Sections II.B and II.C and by a discussion of limitations and transferability in Section III.A. Literature often speaks of data and information without exactly differentiating between these terms. Some authors see in information "refined data", e.g., by placing it in a context of meaning. In this paper, both terms are used synonymously according to the quoted sources.

A. Overview of potential methods for data evaluation

The identification and selection of potential valuation procedures is done through a combination of literature review and in-depth interviews. First, 20 sources of literature are used to collect a comprehensive set of possible valuation approaches. Subsequently, valuation approaches are selected and specified with the help of 50 in-depth interviews with experts from the divisions or departments in the areas of cost engineering, data strategy, data analysis, and purchasing. The consolidated results are shown in Table I [27].

TABLE I. QUALITATIVE AND QUANTITATIVE VALUATION APPROACHES FOR DATA

Method	**Characteristics**			
	Type	*Input*	*Operator*	*Output*
Data Product Scorecard	Qualitative	Data attributes	Scoring-method	Willing-ness to pay
Data Value Design Canvas	Qualitative	Data use case	Expert workshop, Canvas nine factors	Inter-actions / connections
Value determi-nation per user	Quantitative	Acquisition cost Number of users	Discount calculation	Price per user dataset
Value improve-ment by data services	Quantitative	Data material	Statistical analysis, e. g, hypothesis testing	Increase in value or quality through the use of data
Value determi-nation by Laney	Quantitative	Data material	Gartner Valuation Model	Quali-tiative and financial value
Value determi-nation by partners	Quantitative	Theoretical value, maturity, expiration of information	Intangible Assets Evaluation	Monetary information value
Pricing based on customer value	Quantitative	Different data bundels	Versioning, price differen-tiation, surcharge calculation	Price for data bundels

B. Qualitative evaluation

The qualitative assessment and selection of methods is carried out by the procedure described in Section II.A. The following common methods for evaluating vehicle data are identified:

- Data Product Scorecard
- Data Value Design Canvas

The *Data Product Scorecard* is a method of pricing on data marketplaces. For this purpose, the customer's willingness to pay depending on various data properties must first be estimated. This qualitative evaluation of the data properties is made by the Data Product Scorecard from a simulated perspective of end users or potential buyers of the data [21]. As part of an evaluation workshop within the company, the role of the user is taken and each data characteristic given in the scorecard is rated with 0, 5 or 10 points.

The *Data Value Design Canvas* approach looks at the data value chain. The approach is based on the theory of Service Dominant Logic and the "Jobs-To-Be-Done" theory [3][22].

According to [25], the data value chain begins with the generation of data and extends up to the provision of information to the (paying) customer.

C. Quantitative evaluation

Many companies have problems finding the real economic value of their data [23]. For a rethink in the development of new business models [12] and the optimization of internal processes, the determination of this value, especially for the automotive industry, is of particular importance.

Value determination per user

When acquiring companies with data-driven business models who have not yet monetized their database but still offer data-based applications to the end user, the data value is often determined by the value of the application per user. The price of acquisition is divided by the total number of end users of the application. From this calculated price per user the average user acquisition costs are subtracted [16][28].

Value determination by Laney

According to Laney the data is evaluated through quality-based and quantitative financial analysis [20]. In the quality-oriented evaluation, the output is a scoring value between zero and one, in the financial evaluation it is an absolute monetary value. The two-part consideration focuses on methods for improving the "Information Management Discipline" and deals with "Foundational Measures" as:

- How correct, complete and exclusive is the data? (Intrinsic Value),
- How good and relevant is the data for specific purposes? (Business Value),
- How does this data affect key business drivers? (Performance Value).

On the other hand, the "Information Economic Benefit" of "Financial Measures" is examined:

- What would it cost us if we lose this data? (Cost Value),
- What could we get from selling or trading this data? (Market Value),

- How does this data contribute to our bottom line? (Economic Value).

Both considerations provide a quantitatively measurable contribution to the value of data and will be explained in more detail below. Based on a collection and analysis of existing valuation approaches according to Laney, a plausibility check and a requirement analysis lead to new valuation perspectives for the developed new process model (see Section III.A).

Basically, Laney's approaches are not limited to any specific field of application [33]. Section III.A discusses to what extent these approaches are suitable for vehicle data in the automotive industry and in what form adaptations and expansions are necessary for this.

Intrinsic valuation

The intrinsic value of information (IVI) is based on a consideration of information regarding quality and rarity.

$$IVI = Validity * Completeness * (1 - Scarcity) * Lifecycle \quad (1)$$

"Validity" reflects the proportion of "correct" data, "Completeness" the proportion of available data in all potentially accessible data, "Scarcity" the proportion of competitor data to the data available on the market and "Lifecycle" the period of usefulness of the respective information.

Business value of information

When calculating the Business Value of Information (BVI), a process-specific value of data is aggregated for all corresponding processes of the customer's company.

$$BVI = \sum_{p=1}^{n} (Relevance_p) * Validity * Completeness * Timeliness \quad (2)$$

BVI	Business Value of Information
n	Number of considered processes
p	Process index

"Relevance" indicates the usefulness of the data or information for a business process. This value lies in the interval between zero and one. The probability measure "Timeliness" mirrors the aspect that data is not workable at every instant of time.

Performance value of information

The performance value of information (PVI) expresses how the performance of the object of consideration changes with the inclusion of data or information.

$$PVI = \left[\left(\frac{KPI_i}{KPI_c}\right) - 1\right] * \frac{T}{t} \quad (3)$$

PVI	Performance Value of Information
KPI_i	Key Performance Indicator using data
KPI_c	Key Performance Indicator without using data
T	Information lifetime
t	Period of KPI consideration

The key performance indices quantify the power of processes with and without data support. The quotient of both KPIs (KPI_i, KPI_c) conduce to the understanding how data can improve value creation. One example for the "performance value of information" in practice is the data driven Overall Equipment Effectiveness (OEE) for defect prevention [35].

Cost value of information

According to Laney, the cost of information is limited to the process costs for acquiring the data.

$$CVI = \frac{ProcExp * Attrib * T}{t} \left\{ + \sum_{p=0}^{n} LostRevenue_p \right\} \quad (4)$$

CVI	Cost Value of Information
$ProcExp$	Process expenditures
$Attrib$	Proportion of process costs for data acquisition
T	Average information lifetime
t	Period of the process cost measurement
p	Process index
n	Number of processes

The process expenditures "ProcExp" of an overall process are multiplied by a contribution factor "Attrib". The contribution factor expresses which proportion of the process costs can be attributed to the acquisition of the data. "LostRevenue" takes into account the loss of revenue through information shortage in each process.

Market value of information

The market value of information (MVI) is relevant in public marketplaces. The market value of data is expressed, e. g., by a license price (exclusive price) multiplied by the number of licenses.

$$MVI = \frac{Exclusive\ Price * Number\ of\ Licenses}{Premium} \quad (5)$$

In addition to the market price and the number of licenses, the Premium factor takes into account the brand strength and rarity of the data. This Premium factor determines the extra charge a customer would be willing to pay to obtain exclusive rights.

Economic value of information

The Economic Value of Information (EVI) expresses the extent to which revenue changes as soon as the information is used to generate and increase revenue.

$$EVI = \begin{bmatrix} Revenue_i - Revenue_c \\ -(AcqExp + AdmExp + AppExp) \end{bmatrix} * \frac{T}{t} \quad (6)$$

EVI	Economic Value of Information
$Revenue_i$	Revenue with using the information
$Revenue_c$	Revenue without using the information
$AcqExp$	Cost for information gathering
$AdmExp$	Cost of information administration
$AppExp$	Cost for the information application
T	Information lifetime
t	Period of the measurement

First, the revenue is recorded without the use of certain relevant information. Subsequently, the turnover is estimated by a control group using this information. Also the expenditures for obtaining, storing, managing, and using information are identified. The respective revenues and expenditures are measured in a period t. The economic value of the information is now calculated by subtracting the revenues without information and the costs incurred from the revenues gained with the information. The resulting value is multiplied by the ratio of information lifetime T to duration of measurement t. This results in an EVI value for the full information lifecycle.

III. Developed Methodology

A. Motivation of a new valuation approach

IVI, BVI and PVI are pure comparative methods. These do not calculate monetary values, yet they show some interesting perspectives. The BVI sums up the value of special information over all use cases in the company, and both the rarity and the quality indicators are taken into account in the IVI.

The methodology introduced in this section follows an approach also based on PVI and BVI. The value of data can lead to process improvements and profit increases. The Data Value Design Canvas lists nine factors which affect the value of data. The effects of information/data are considered generally. For example, information/data protects against unwanted events or promotes wanted events. Unwanted events always result in costs. Thus, the avoidance of unwanted events corresponds to a cost reduction. The realization of desired events effects an increase in sales as the most relevant example.

The Data Product Scorecard assesses the willingness to pay of a customer. If the information considered is "perfect", the customer is willing to pay the full price for this information. The determination of willingness to buy is based exclusively on estimates.

The method according to Laney, the Data Value Design Canvas and the Data Product Scorecard focuses on the quality of the information/data.

The monetary impact of EVI relates to an increase in sales. A monetary value can be derived from the Data Product Scorecard by multiplying the qualitative result score by the willingness to pay. A market value, in turn, may result from a fixed license price multiplied by the number of licenses sold or salable and a rarity and reputation factor.

All methods have in common that always certain use cases of data usage are considered. The PVI and BVI consider process improvements, the MVI and the Data Product Scorecard data sales, the Data Value Design Canvas both data sales and process improvements. Only the IVI does not aim at a defined application field because it is based on a general quality criterion.

By comparing the different approaches listed in Section II, some requirements can be derived for a new concept integrating different aspects:

- Quality factors must be taken into account.
- The willingness to pay is relevant for data sales. This depends on various factors, including the purchase motive, the perceived benefit, the reputation of the seller and the individual purchase situation.
- Data has the potential to increase sales on the customer side or to reduce costs for internal customer processes.
- Competition should be considered as an important factor.

B. Evaluation model

The developed "integrated methodology" for an innovative evaluation model meets these requirements by a combination of quality assessment [26], price differentiation [36], cost management, and competitive analysis [32]. So, for a specific use case, it is possible to estimate a monetary value of data by integrating the various value perspectives. The plausibility of selling prices is achieved by combining qualitative tools based on methods such as Business Model Canvas or Data Canvas with a practical evaluation through quality workshops as well as quantitative calculations. These include, among other things, the valuation by Laney, a bottom-up cost calculation as well as profit split approaches.

Figure 1 shows the process steps of the model for a use-case-specific value determination. The non-rivalry property of data enables multiple sales of similar data bundles or even the same dataset. The total value of the data bundle can be determined as the sum of the values across all (potential) use cases:

$$V_G = \sum_{i=1}^{n} V_i \quad (7)$$

V_G	Total value of data bundle G
V_i	Individual data value for the use case i
i	Use case index
n	Number of use cases

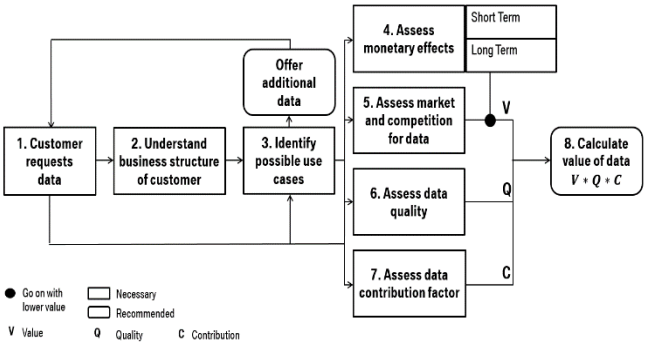


Figure 1. Process for the monetary valuation of data

As a first step in the process, data requests from potential business customers are collected or data is proactively offered to potential customers. In order to be able to identify potential data needs, a customer's business model must first be understood (1 + 2).

In which way the customer translates the data into benefits can be identified through a systematic analysis of possible use cases. For this a combination of the Business Model Canvas and Data Value Design Canvas is suitable. With progressive understanding of the application, it is theoretically possible to offer the customer additional versions of data bundles (2 + 3).

In order to determine the customer's willingness to pay, the value of the use case must be understood in detail [15]. For this, a possible cost reduction or increase in sales by the data must be determined.

For each use case, there is both a short-term and a long-term monetary benefit, which in individual cases can also be zero. The model of Figure 1 shows a parallel approach to Laney's business value calculation of information, which determines the data relevance for specific processes (4).

In cases where there is competition on "data marketplaces" or the self-collection of data is significantly more favorable than granting the monetary benefit to a third party, these influences must be measured for further calculation (5).

At the same time, the data is qualitatively evaluated based on selected criteria (see Section II.B). The model considers the monetary value of quality criteria, following Laney's valuation ideas in the form of a quality factor Q (see Section IV.D) (6).

In addition, the data contribution factor C takes into account that other vehicle generated data or additional information may be necessary in addition to the offered vehicle usage data (see Section IV.D) (7).

After having carried out steps 1 to 7, finally the use-case-specific combined value of the data is calculated. The determined preliminary data value V (see Section IV.B) is multiplied by the quality factor Q and the data contribution factor C. Both of them are discount factors. The preliminary value and the willingness to pay is reduced according to low quality or insufficient amount of data. Interdependencies between the factors are possible but not taken into account.

C. Restrictions

It is important to make the best selection of use cases to generate the maximum revenue of data sales for the business.

$$R = \sum_{i=1}^{n} RUC_i \tag{8}$$

R	Aggregated revenue
RUC_i	Revenue of use case i
i	Index of implemented use cases
n	Number of implemented use cases

The present technical restrictions must be observed. However, this limits the number of possible applications. The possible amount of data to be transferred to different use cases is restricted by the number of transmission paths and the transfer rate of each of them:

$$TL_k \geq \sum_{j=1}^{m} D_{jk} \tag{9}$$

TL_k	Maximum transfer rate for transmission k
D_{jk}	Data volume j to be transferred by path k
k	Index of possible transmission paths
j	Index of required data for each transmission path depending on the requirements of the application
m	Total number of required data volumes per transmission path

Since many use cases are based on fleet data, the complexity is additionally increased. Accordingly, it has to be determined which vehicles transmit which data via which path with which transfer rates, in order to technically enable the optimum amount of use cases. This problem cannot be solved conventionally or manually. It is necessary to test new tools for this, such as machine learning algorithms.

IV. Use Case Results

The applicability of the developed methodology is experimentally tested on the basis of three real sales situations. The first use case deals with the sale of weather data to a weather service provider who wants to offer additional hazard warning services for autonomous driving.

The second use case relates to the sale of weather data to a transmission system operator to “ensure a reliable and uninterrupted supply in the high voltage grid for approximately 41 million people” [34].

In the third use case the sale of road segment data (RSD) to a navigation maps provider is considered. This data helps

to provide a high definition road map for autonomous driving.

A. Business structures of the use cases

Table II summarizes the results of the structural analysis of the business models. The findings serve as a first qualitative evaluation of the business structure, from which the data needs of the customers and the data bundles offered can be derived.

TABLE II. BUSINESS STRUCTURE ANALYSIS [17][18][29][34]

Object of investigation	**Customer**		
	Weather service	***Transmission system operator***	***Navigation maps provider***
Business Structure	***Value Proposition*** Deliver weather service ***Key resources*** Internal rational data: - External rational data: Satellite pictures Internal continuous data - External continuous data: Weather station information, weather car data ***Key Activities*** Gather and refine data for weather forecast ***Customers*** Companies, end users ***Segments*** Automotive, governments, software companies ***Relationships*** Direct ***Channels*** Direct contact	***Value Proposition*** Secure reliable supply of electricity ***Key resources*** Internal rational data: Electricity prices External rational data: Consumer behavioral studies Internal continuous data Electricity supply, electricity demand External continuous data: Weather station data, weather car data ***Key Activities*** Transportation of electricity, maintaining energy balance ***Customers*** Companies ***Segments*** Large industries, consumers, governments ***Relationships*** Indirect	***Value Proposition*** Build tomorrows road network ***Key resources*** Internal rational data: Street map data External rational data: - Internal continuous data Navigational data, destination External continuous data: Real time traffic information, car data (RTTI), Road segment data (RSD) ***Key Activities*** Gather and refine map data ***Customers*** Companies ***Segments*** Automotive industries ***Relationships*** Direct ***Channels*** Direct contact ***Revenue Stream*** Selling refined information (HD map) for autonomous driving, location based services ***Cost Structure***

Object of investigation	**Customer**		
	Weather service	***Transmission system operator***	***Navigation maps provider***
	Revenue Stream: Selling refined weather forecast information ***Cost Structure*** Data buying, personnel	contact ***Channels*** Customer events, customer committees ***Revenue Stream*** Offering connection and transmission services, maintenance of energy balance, operation of energy exchange, offshore balancing ***Cost Structure*** Grid connection, transmission services, maintenance of energy balance	Personnel, data transfer (automotive industry)
Data offering	Sensor data (rainfall, road surface texture)	Temperature, GPS (latitude, longitude)	RSD (camera data: edge markings, center markings, strip width, crash barriers, guide posts, signs, wild warning reflectors and barriers)
Possible Use Case	Local Hazard Service	Solar Energy Prognosis	HD-Map

In the business structure analysis, data is divided into four categories. “Internal rational data” is owned by the organization and updated at certain points of time. “Internal continuous data” is owned by the organization and available in the form of continuous streams. “External rational data” is owned by third parties and updated at certain points of time. “External continuous data” is owned by third parties and available as continuous streams. This classification is used for the data contribution factor defined in Section IV.D.

B. Monetary Effect

Next, the financial perspective is taken and the monetary effect of the three use cases is considered.

Weather service

Table III shows a forecast for the sales figures for fully and partially automated vehicles [7].

TABLE III. SALES FORECAST FOR (PARTIALLY) AUTONOMOUS VEHICLES

Year	2014	2020	2035
Sales figures in millions	0.8	3.3	28

With an assumed willingness to pay for the Hazard Warning Service of 10 € per (partially) autonomous vehicle per year and the estimated globally available 3.3 million high or fully autonomous vehicles in the year 2020, this results in a potential revenue of 4.95 million €. The calculation assumes a market share of 15%:

$$3.3\ Mio * 10€ * 0.15 = 4.95\ Mio\ €$$

Market shares of up to 25% are often predicted in this product category [36]. In this best-case scenario, the weather service provider reaches a market share of 3.3 Mio * 0.25. This number is based on market data for navigation charts [36]. In this case, the data buyer reaches a revenue of 8.25 million €.

$$3.3\ Mio * 10€ * 0.25 = 8.25\ Mio\ €$$

For the further demonstration of the methodology the lower potential revenue of 4.95 million € is assumed.

Transmission system operator

The data sets offered in this use case are generated by a dynamic car-sharing fleet of 650 vehicles, covering the entire business area of Munich (82 km^2). The value of the data is analyzed on the one hand by means of a pure cost estimate and on the other hand by an opportunity analysis. In total, the transmission system operator allocates 1 billion € to grid stabilization in Germany in 2017 [18][34]. According to Statista, the total turnover of the German network operators in 2015 was 31.2 billion €, in which the company in question realized 3.3 billion [24][34]. This results in a market share of 10.6%.

$$\frac{3.3\ Bio\ €}{31.2\ Bio\ €} = 10.6\%$$

In 2017, 547 TWh of electricity were produced in Germany [9]. Adjusted for the market share of 10.6%, 57.8 TWh of electricity are attributable to the company.

$$547\ TWh * 10.6\% = 57.8\ TWh$$

Since the company also has another location, the total transfer performance must be considered. At the second location, the company is the sole network operator [34]. There are about 106 TWh of electricity generated [11]. This results in a total transmission capacity of 153.8 TWh per year.

$$57.8\ TWh + 106\ TWh = 153.8\ TWh$$

The necessary data for the solar energy forecast (temperature and GPS coordinates) resulting from business model analysis are delivered via the dynamic vehicle fleet in the Munich area. The power consumption in Munich is identified as a reference. In Munich, 2.8 TWh of electricity are consumed per year in total of all inhabitants [30][37]. Compared to the total transfer performance of the company, this results in a share of 1.7% for the Munich area.

$$\frac{2.8\ TWh}{153.8\ TWh} = 1.7\%$$

If one multiplies this percentage with the amount of one billion €, which the network operator identifies as costs for network stabilization in 2017 [34], this results in a maximum possible cost reduction of 17 million euros, which the network operator saves in a best-case scenario, in which perfect network stabilization over the data supply is provided.

$$1\ Bio\ € * 1.7\% = 17\ Mio\ €$$

However, it can be assumed that the provided data cannot be expected to improve the forecast of renewable energy performance for grid stabilization perfectly. Nevertheless, even the assumption of a ten percent improvement leads to a cost reduction of 1.7 million €.

This achievable value is calculated taking into account the average solar energy generated in Bavaria. The value of the data material is justified by consideration of an opportunity value. Here, the value of data for a perfect solar energy forecast results from the price and the amount of solar energy in the Munich area, which would have to be bought in the absence of such a data-based demand forecast. In other words, the amount “price times quantity” reflects the short-term purchase of solar energy which is necessary for grid stabilization, resulting from inadequate predictions. This could be avoided by making good use of the data. The Fraunhofer Institute gives an average value of produced solar energy of 9.8% in electricity generation in Germany [13]. The Federal Network Agency's statistics show that Bavaria produces the most solar energy in Germany [8]. Munich is the leader in hours of sunshine [19], which is why for Munich the average value of produced solar energy is calculated as the lower limit in a minimum scenario.

$$2.8\ Twh * 9.8\% = 0.27\ TWh$$

Taking into account a green electricity price of 56 € per MWh, this results in a potential saving of 14.4 million € (assuming a perfect prediction) [10].

It is striking that the calculated values of the best possible savings are relatively close (14.4 million € and 17 million €).

Navigation maps provider

The calculation for this use case is similar to the monetary revenue estimate for the weather service. The willingness to pay for the HD card in the automotive industry is 60 € for a highly autonomous or fully autonomous vehicle per year.

For the year 2020, a global volume of 3.3 million high or fully autonomous vehicles is expected on the market, in 2035 a total fleet of 28 million vehicles (see Table III).

The market for navigation charts in vehicles is divided into market shares of 15% to 25% [36].

These assumptions lead, in the worst case scenario (assuming a market share of 15%), to a potential turnover of at least 29.7 million € in 2020 (3.3 million * 60 €* 0.15) and 252 million € in 2035 (28 million * 60 € * 0.15).

C. Market competition

The market position is qualitatively described for all three use cases by the criteria according to [23] "is it valuable", "is it rare", "is it hard to imitate" and "is the firm organized for success". The answers to the questions are based on a competitive analysis in which 22 competitors are considered in the case of the transmission system operator, 24 for the weather service and 26 for the navigation maps supplier. The qualitative assessment shows that there is a "short term competitive advantage" for all three use cases. In the case of the transmission system operator, also a quantitative value can be set by an "in-house manufacturing vs. external purchase analysis", which means a counter-calculation of an internal generation of data as an alternative to an external purchase. From practical experience, a rigid sensor system of 2000 sensors for the price of 145 € per sensor is assumed. Surcharges are 8,000 € for housing development, 90,000 € for programming and 5,000 € for position planning.

This results in investment costs of 393,000 euros:

$$2000 * 145\,€ + 8{,}000\,€ + 90{,}000\,€ + 5{,}000\,€ = 393{,}000\,€$$

Furthermore, a monthly failure rate of 0.5% is assumed. This results from damage done by weather and vandalism to publicly accessible sensors. Corresponding monthly repair costs are 1,850 €. For data transmission 20 € per month are estimated. This results in a total annual expenditure of 415,440 €:

$$393{,}000\,€ + 12 * 1{,}850\,€ + 12 * 20\,€ = 415{,}440\,€$$

This consideration is a minimum estimate. It can be assumed that the transmission system operator does not own 2,000 relevant properties in Munich to set up sensors.

The value of 415,440 € corresponds to the amount that the transmission system operator would have to raise to collect perfect data in the same quantity.

D. Data quality and contribution factors

In order to evaluate the data quality, 14 data scientists are interviewed in four expert workshops for each case on given quality criteria. The results of the workshops are summarized in Figure 2 and transformed into a quality factor Q.

$$Q = \sum_{i=1}^{n} c_i * w_i \qquad (10)$$

Q Quality factor
c_i Evaluation factor of criterion i
w_i Weight of criterion i
i Criterion index
n Number of criteria

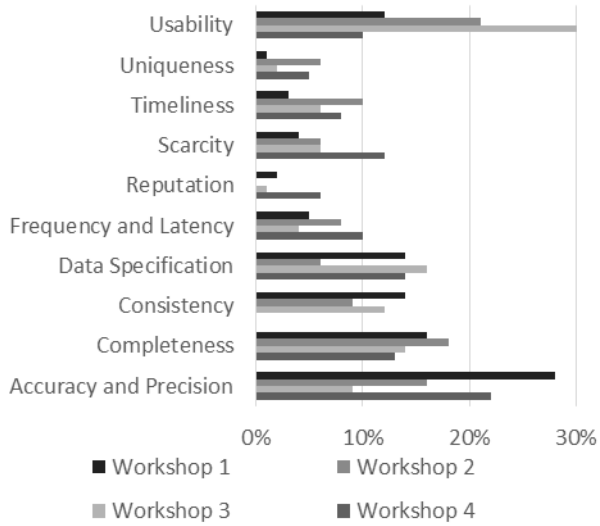


Figure 2. Evaluation of the quality criteria of data

The evaluation factor values c_i of the criteria result from a pairwise comparison of all criteria in a preference matrix. The data contribution factor *C* expresses if all required data (contribution factor = 1), almost all data (contribution factor = 0.75), about half of the data (contribution factor = 0.5), few data (contribution factor = 0.25) or no data at all (contribution factor = 0) can be provided.

$$C = CM * w_{cm} + CK * w_{ck} \qquad (11)$$

C Data contributing factor
CM Contribution factor of metadata
CK Contribution factor of key data
w_{cm} Evaluation of metadata
w_{ck} Evaluation of key data

The use-case-specific factors CM und CK differentiate between key data und metadata (additional data). The expert based weights w_{cm} and w_{ck} rate the relative data contribution of each data type in the use case. Location based services are one example. They are based on customer preferences (metadata) on the one hand and GPS data (key data) generated by the vehicle on the other hand. Experts give weights of 0.2 for GPS data and 0.8 for preference data. It is assumed that the vehicle can deliver 90 percent of the GPS data needed, however none of the preference data. So, the contribution factor of the metadata is zero. The calculation of C results in the value of 0.18:

$$0.2 * 0.9 + 0.8 * 0 = 0.18$$

The corresponding calculation rules provide the following results (see Table IV) coming from the workshops for the three use cases.

TABLE IV. WORKSHOP RESULTS

Key figure	Quality factor	Contribution factor
Weather service	0.79	0.5
Transmission system operator	0.8	1.0
Navigation maps provider	0.8	0.55

E. Selling prices

Through the product of monetary effect (*V*), quality (*Q*) and contribution factor (*C*) (see Figure 1) a value for the offered data of 1.95 million € (4.95 million € * 0.79 * 0.5) for the use case "weather service" is calculated, for the use case "transmission system operator" 332,352 € (415,440 € * 0.8 * 1.0) and for the use case "navigation maps manufacturer" in the worst case scenario 13.07 million € (29.7 million € * 0.8 * 0.55). With the weather service as well as the navigation maps manufacturer the turnover is in the foreground. In the case of the transmission system operator, the focus is not on turnover, but on cost savings. The costs for the self-collection of the data are less than the calculated costs of externally purchased data that lead to the same savings effect, so that this lower value of about 415,440 € is used. In the case of self-collection of data the contribution factor is 1.0, since all required data is recorded internally.

F. Lessons Learned

On the one hand the presented evaluation model was applied to several ficitious use cases with real information coming from companies interested in buying data but without any decision. On the other hand, the evaluation model was tested on several specific sales situations and the outcome of the model, i.e., the post calculated data value, was compared to the real sales price. The monetary data values determined using the evaluation model show an average deviation of 8% from the sales prices negotiated in practice.

V. CONCLUSIONS

The methods of data evaluation identified in the literature are individually not suitable for practical value determination of data and their pricing in sales situations. This paper presents a methodology that focuses on the selling of data as intangible products to external business partners.

The methodology can also be transferred to use cases within the company. In addition to determining the value of the data, decisions regarding the pricing model must be made.

However, for long-term strategies it is unclear to what extent recorded data is valuable in the future. Data that is still useless, because currently there are no use cases, can be highly relevant for future use cases. Because of the existing knowledge gap and missing empirical values, it is impossible to determine a value of data over the entire lifecycle, above all because of very uncertain future potentials.

This article exemplifies a possible evaluation and monetization of a small fraction of the total data available in the automotive industry.

Looking at the huge amounts of data available there, it quickly becomes clear that due to technical limitations probably never all potential use cases can be implemented. There are various transmission options for vehicle generated data. The built-in memory can be read in authorized garages, updates can be transmitted at weekly or daily intervals, or data can be transferred in real time. There is always a technical limitation due to the restricted transfer rates or transfer options. Not all conceivable applications can be realized at the same time.

It is also an open question whether it makes sense to regard the vehicle as an open platform. In this case, an automobile manufacturer or even the automotive industry as a platform provider could probably sell the platform as a service (PaaS) to service providers who will pay for specific data accessed via the platform. As an analogy, platforms of Apple and Android can be considered. Third parties develop services to be offered on these platforms. The developed services (e. g., apps) increase the attractiveness of the platform. Depending on the design, there are direct and indirect network effects. With regard to autonomous driving, this approach may potentially increase the attractiveness of vehicles and vehicle fleets acting as such platforms. For example, at BMW, in addition to many existing Connected Drive services, applications of third-party providers can be activated, which leads to an immense increase of the value of a ride and the driving experience for the customer. Here, completely new service ecosystems spanning and connecting different industrial sectors are appearing. To name only one step toward the future, the intelligent personal assistant from BOSCH enables the networking of car services and e-home services [6].

REFERENCES

[1] F. Bodendorf, T. Meissner and J. Franke, "Valuating and Pricing of Vehicle Generated Data as a Marketable Product in the Automotive Industry," The Eighth International Conference on Advances in Vehicular Systems, Technologies and Applications (VEHICULAR 2019), Jun. 2019, pp. 16-21, ISSN: 2327-2058, ISBN: 978-1-61208-720-7

[2] J. Balasubramanian et al., "Car data: paving the way to value creating mobility : Perspectives on a new automotive business model," Advanced Industries McKinsey, 2016

[3] L. A. Bettencourt and A. W. Ulwick, "The customer-centered innovation map," Harvard Business Review, vol. 86, pp. 109-114, 2008

[4] BMW. *BMW ConnectedDrive Customer portal - digital networking to your BMW.* [Online]. Available from: https://www.bmw-connecteddrive.de/app/index.html#/portal/store/Base_TolCarOffer. Retrieved [05-31-2020].

[5] BMW. *NOW mobility services* . [Online]. Available from: https://www.bmwgroup.com/de/marken/now-mobilitaetsdienstleistungen.html. Retrieved [05-31-2020].

[6] BOSCH. *Only driving was yesterday - the personal assistant is tomorrow*. [Online]. Available from: https://www.bosch-presse.de/pressportal/de/de/nur-fahren-war-gestern-%E2%80%93-der-persoenliche-assistent-ist-morgen-101568.html. Retrieved [05-31-2020].

[7] A. Brugger. *Global production forecast for semi- and fully automated vehicles*. [Online]. Available from: https://www.oliverwyman.com/con-tent/dam/oliver-wyman/global/en/files/who-we-are/press-releases/OliverWyman_Graphics_Value%20Pools%20Autonomous%20Driving_EN_16072015_final.pdf. Retrieved [05-31-2020].

[8] Bundesnetzagentur. *Figures, data and information on the EEG*. [Online]. Available from: https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/ErneuerbareEnergien/ZahlenDatenInformationen/zahlenunddaten-node.html. Retrieved [05-31-2020].

[9] B. Burger. *Electricity generation in Germany in 2017*. [Online]. Available from: https://www.ise.fraunhofer.de/content/dam/ise/en/documents/News/Stromerzeugung_2018_2_en.pdf. Retrieved [05-31-2020].

[10] e-control. *Current market price pursuant to § 41 Green Electricity Act*. [Online]. Available from: https://www.e-control.at/en/industrie/oekoenergie/oekostrommarkt/marktpreise-gem-paragraph-20. Retrieved [05-31-2020].

[11] L. Eglitis. *Energy balance in the Netherlands.* [Online]. Available from: https://www.ebn.nl/en/the-energy-balance-in-the-netherlands/. Retrieved [05-31-2020].

[12] J. L. L. Francisco and J. Esteves, "Value in a Digital World: How to assess business models and measure value in a digital world," Springer, 2017, ISBN 978-3-319-51750-6

[13] Fraunhofer. *Share of renewable energies | Energy Charts*. [Online]. Available from: https://www.energy-charts.de/ren_share_de.htm?year=2018&source=solar-share&period=monthly. Retrieved [05-31-2020].

[14] W. Gründinger. *Data-Driven Business Models in Connected Cars, Mobility Services & Beyond*. [Online]. Available from: https://www.bvdw.org/fileadmin/bvdw/upload/publikationen/connected_mobility/20180418_data_driven_business_models_Seiberth_Gruendinger.pdf. Retrieved [05-31-2020].

[15] R. Harmon, H. Demirkan, B. Hefley and N. Auseklis, "Pricing Strategies for Information Technology Services: A Value-Based Approach" 42nd annual Hawai'i International Conference on System Sciences (HICSS 2009), IEEE Press, Jan. 2009, pp. 1–10, ISBN: 978-1-424-44197-6

[16] N. Henke et al., "The Age Of Analytics: Competing In A Data-Driven World," McKinsey Global Institute, 2016

[17] HERE. *Location-Based Services for the Autonomous Future.* [Online]. Available from: https://www.here.com/en. Retrieved [05-31-2020].

[18] M. Holland. *Power grid under pressure: Tennet reports record costs for emergency interventions*. [Online]. Available from: https://www.heise.de/newsticker/meldung/Stromnetz-unter-Druck-Tennet-meldet-Rekordkosten-fuer-Noteingriffe-3929093.html. Retrieved [05-31-2020].

[19] M. Kniekamp, Spitzenreiter. *In this statistic Munich is Germany's showcase city*. [Online]. Available from: https://www.merkur.de/bayern/muenchen-ist-laut-studie-zdf-deut-sche-stadt-mit-meisten-sonnenstunden-9855027.html. Retrieved [05-31-2020].

[20] D. B. Laney, "Infonomics: How to monetize, manage, and measure information as an asset for competitive advantage," Routledge, 2018, ISBN: 978-1138-09038-5

[21] E. Leonidas, DataStream. *A Practical Guide to Pricing Data Products.* [Online]. Available from: https://cdn2.hubspot.net/hubfs/573334/Downloadable_Content_(WP_or_Guides)/DataStreamX_Data_Product_Pricing_Whitepaper.pdf. Retrieved [05-31-2020].

[22] C. Lim et al., "From data to value: A nine-factor framework for data-based value creation in infor-mation-intensive services," International Journal of Information Management, vol. 39, pp. 121–135, 2018, ISSN: 0268-4012

[23] S.M. Liozu, W. Ulaga, "Monetizing Data, A Practical Roadmap for Framing, Pricing & Selling Your B2B Digital Offers," VIA Publishing, 2018, ISBN: 978-1-945815-04-1

[24] D. Loesche. *Infografik: Comparison of the four German transmission system operators*. [Online]. Available from: https://de.statista.com/infografik/6023/vergleich-der-vier-deutschen-eebertragungsnetzbetreiber/. Retrieved [05-31-2020].

[25] K. Mathis and F. Köhler, "Data-Need Fit - Towards data-driven business model innovation" The Fifth Service Design and Innovation conference (ServDes.), pp. 458–467, 2016, ISBN: 978-91-7685-738-0

[26] D. Mcglivary, "Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information: Definitions of Data Categories" Chief Data Officer & Information Quality Symposium, 2009

[27] T. Meißner, "Value and utilization of data on hardware and software usage in the vehicle," University of Erlangen-Nuremberg (Master Thesis), 2018

[28] K. O'Neal, "Quantifying the Value of Data: The First Step in Data," unpublished presentation

[29] A. Osterwalder and Y. Pigneur, "Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers, Hoboken," John Wiley & Sons, 2010, ISBN: 978-0470-87641-1

[30] Photovoltaik Rechner. *Information, Calculation & Companies*. [Online]. Available from: https://www.rechnerphotovoltaik.de/photovoltaik/in/bayern/muenchen. Retrieved [05-31-2020].

[31] B. Schmarzo and M. Sidaoui. *Applying Economic Concepts To Big Data To Determine The Financial Value Of The Organization's Data And Analytics, And Understanding The Ramifications On The Organizations' Financial State-ments And IT Operations And Business Strategies*. [Online]. Available from: https://infocus.delltechnologies.com/wp-content/uploads/2017/04/USF_The_Economics_of_Data_and_Analytics-Final3.pdf. Retrieved [05-31-2020].

[32] C. Shapiro and H. R. Varian, Versioning, "The Smart Way to Sell Information," Harvard Business Review, vol. 6, pp. 107 f., 1998

[33] F. Stahl and G. Vossen, "High quality information provisioning and data pricing" 29th International Conference on Data Engineering Workshops (ICDEW 2013), 2013, pp. 290-293

[34] TenneT Holding B.V.. *Annual report 2017 TenneT Holding B.V*. [Online]. Available from: https://annualreport.tennet.eu/2017/annualreport. Retrieved [05-31-2020].

[35] J. H. Thun, "Maintaining preventive maintenance and maintenance prevention: analysing the dynamic implications of Total Productive Maintenance," System Dynamics Review: The Journal of the System Dynamics Society, vol. 22, pp. 163-179, 2006

[36] S. Viswanathan and G. Anandalingam, "Pricing strategies for information goods," Sadhana Indian Academic Sciences Journal, vol. 30, pp. 257–274, 2005, ISSN: 0256-2499

[37] Wirtschaftswoche. *GPS-Satellitennavigation: Wie Navigationsgeräte-Firmen um einen Milliarden-Markt kämpfen.* [Online]. Available from: https://www.wiwo.de/unternehmen/gps-satellitennavigation-wie-navigationsgeraete-firmen-um-einen-milliarden-markt-kaempfen-seite-3/5426542-3.html. Retrieved [05-31-2020].

Two Models for Hard Braking Vehicles and Collision Avoiding Trajectories

Fynn Terhar

BMW Group / FernUniversität in Hagen
Department for Fleet Intelligence
Munich, Germany
Email: fynn.terhar@bmw.de

Christian Icking

FernUniversität in Hagen
Department for Cooperative Systems
Hagen, Germany
Email: christian.icking@fernuni-hagen.de

***Abstract*—In this paper, we describe two models that describe vehicle dynamics in full braking situations with collision avoiding motions. By combining the equations of the classic Ackermann-Model with conditions that ensure a stable vehicle movement during simultaneous heavy braking and turning motions, we derive two models that describe the set of controllable trajectories by compound equations in the x, y plane. We describe a simplified model first and compare its performance to the well known Constant-Turn-Rate-And-Acceleration-Model, which is computationally more expensive and less precise. We discuss the simplified model regarding uncertainties and their effect on reachability estimation of vehicles in admissible scenarios, to show the feasibility of our solution. By considering uncertainties of the parameters used in the Basic Model, we show a way to estimate the reachable area of a hard braking vehicle in different starting constellations. We extend this Basic Model to handle much more dynamic situations and starting conditions. In the new Extended Model, the initial yaw rate is an input, as well as the maximum steering angle change rate. The vehicle length and steering direction are also considered. By these additions, we are able to describe trajectories in much more realistic detail than with the Basic Model. We derive and present all necessary equations required for computing these trajectories. Furthermore, we analyze and demonstrate all possible types of trajectories that directly follow from our definitions.**

Keywords–Reachability; Trajectory; Dynamic Vehicle Model; Safety; Collision Avoidance; Braking; Trajectory Types.

I. Introduction

Many functions in Highly Automated Driving (HAD) and Advanced Driving Assistance Systems (ADAS) are discussed regarding their safety towards events caused by other traffic participants, whose behavior is not well predictable. In case of an unforeseen event, vehicles need to avoid a collision by a suitable trajectory. In literature, these trajectories are often referred to as Fail-Safe-Trajectories. These trajectories can either be evasive and try to find a solution around an obstacle or bring the vehicle to an emergency stop. The vehicle is then forced to find a trajectory till full stop within an area in front of the vehicle, which is defined by its physical properties and speed vector.

A. Motivation

In this paper, we call this area the *Braking Area*, which is important to know in many different applications and situations, such as in Figure 1, in which two vehicles are unexpectedly confronted with each other. Also, when defining the set up of on-board sensors, it can be useful to have a good knowledge of the braking area. Another example is the search for fail-safe trajectories, such as shown in Figure 2, where the knowledge of the reachable set of vehicle states can significantly accelerate the computation, as it reduces the search space and can therefore save valuable time in emergency situations.

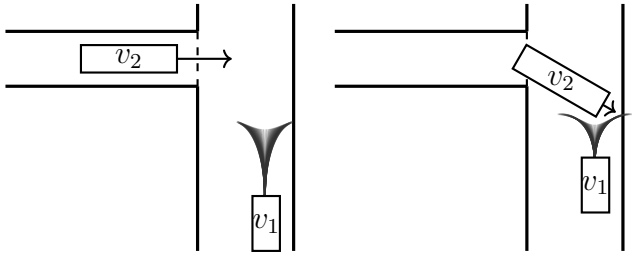


Figure 1. Left, two vehicles v_1, v_2 approach a crossing. Right, at sudden confrontation, v_1 can benefit from its reachable area for emergency braking.

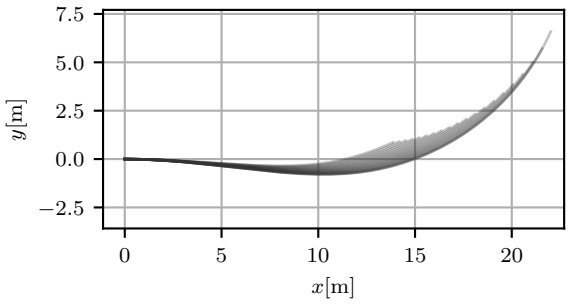


Figure 2. A set of 20 braking trajectories computed by the Extended Model introduced in this paper. Advanced dynamics at the start state are incorporated.

B. Literature overview

This paper is based on our previous work introduced in [1], where we introduce a new model to compute trajectories of hard braking and collision avoiding vehicles. This task is related to finding fail-safe trajectories. Methods for avoiding obstacles are numerous, see for example Werling et al. [2], where the authors address dynamic street scenarios by an optimal control approach. The method generates trajectories that are optimal in terms of jerk minimization and following a previously computed trajectory. Another approach is explained by Ziegler et al. [3]. They use a cost function to plan obstacle avoiding paths in unstructured environments, but not on the

description of fail-safe trajectories. Several approaches towards finding fail safe trajectories for road vehicles exist. Pek and Althoff [4] describe a method to generate fail-safe trajectories for dynamic traffic scenarios in a computationally efficient manner. Their solution approximates the set of reachable states of the ego vehicle and other traffic participants and can therefore guarantee collision free trajectories. A motion planner for fail-safe trajectories is shown by Magdici and Althoff [5]. A related application is presented in [6], where a safety framework is demonstrated that can test a planned trajectory for possible future collisions. A complete motion planning system is described by Heinrich [7]. The presented sampling based approach consists of three cyclic, elementary steps. State Space Exploration uses vehicle surround view sensors. Based on the explored state space, trajectory samples are generated during Trajectory Generation phase. The last step consists in Optimization, which means finding an optimal trajectory from the previously generated trajectories.

Mitchell et al. [8] discuss different approaches of reachability analysis of dynamic systems for the safety assessment of trajectories. Asarin et al. [9] present an approach for reachability approximation of partially linearized systems in general. An often applied technique to approximate the state space efficiently is by zonotopes, see, e.g., the paper of Girad [10]. Koschi et al. [11] introduce an open source software solution which predicts road occupancy by traffic participants within a given time horizon. By overestimating the occupancy by the union of several object models, the authors ensure to find all possible traffic configurations. Potential braking and turning is overestimated by a circle of lateral and longitudinal maximum and minimum accelerations. The physical interaction between velocity and admissible lateral accelerations are therefore overestimated. Althoff [12] describes many underlying concepts of reachability analysis for road vehicles. In contrast to formal verification, ByeoungDo et al. [13] propose a Recurrent Neural Net for predicting traffic participants. Explicit braking and turning motions and their interrelation are not in the focus. Both of our models provide a more detailed and accurate description of this interaction in order to reduce the overestimation towards a more realistic model.

The interrelation of braking and turning is, e.g., discussed by Giovannini et al. [14] where the authors describe the last point in time when a collision can be avoided by swerving. The authors explicitly focus their work on two-wheeled vehicles. Ackermann et al.[15] present control strategies for braking and swerving motions. Choi et al. [16] propose an additional strategy based on model predictive control.

C. Contribution

In this paper, we present two different models for calculating feasible trajectories for braking while turning. In the first part, we describe a Basic Model that can quickly calculate motion primitives for estimating the reachable area of a vehicle while simultaneously braking and turning, further on called *Braking Area*. This *Basic Model* is simplified in two essential points. Firstly, the yaw rate at time $t = 0$, which is the change in direction a vehicle is pointing, was implicitly assumed to be zero. Secondly, the rate at which the yaw rate $\dot{\psi}$ can change, was assumed to be ∞. In order to estimate the Braking Area, these assumptions are valid as they include all trajectories with more realistic properties, and hence lead to an acceptable overestimation of the breaking area. However, when calculating feasible braking trajectories in scenarios that deviate much from the aforementioned assumptions, a more realistic model is to be preferred. Therefore, we extend the Basic Model towards an *Extended Model* by including the yaw rate at time $t = 0$ as $\dot{\psi}_0$, as well as a new model parameter $\dot{\hat{\delta}}$, which limits the change rate of $\dot{\psi}$ by changing the steering angle δ. See Figure 2 for an example in which a set of Extended Model trajectories are shown. Furthermore, we extend the Basic Model by the turning direction s, which determines the sign of a trajectory's curvature. With these extensions, we drastically raise the applicability towards being able to:

1) Calculate realistic trajectories, also for dynamic situations at $t = 0$.
2) Call the model recursively, as all produced outputs my be fed back as inputs.
3) Use the model to calculate motion primitives for any start state of moving vehicles which is in a stable state, e.g., is not sliding over the road uncontrollably.
4) Search for feasible and yet complex emergency trajectories by concatenating motion primitives.

With these two models, we contribute equations that can be used to calculate trajectories for automated vehicles or to estimate the reachable area in emergency braking situations. With respect to the work of Heinrich [7], our equations contribute trajectories for the Trajectory Generation step of vehicle motion planning for automated driving.

In Section II, we describe underlying assumptions that hold for both models and outline the major differences. Used symbols and notations are also described. In Section III we derive the Basic Model, which directly calculates vehicle trajectories towards a full stop while simultaneously braking and steering under simplified assumptions. Section III-A introduces all equations of the Basic Model. Braking and steering always needs to be performed in a balanced way, as both influence the controllability of the vehicle on the road. We therefore introduce a parameter that describes the ratio of this compromise. Furthermore, the friction between different road surfaces and tires is considered. A comparison of the introduced Basic Model and the CTRA-Model is given in Section III-B. The influence that uncertain model inputs and parameters have are discussed in Section III-C. The Extended Model is introduced in Section IV. The model equations of the Extended Model are derived in Section IV-A, which takes the change rate of steering angle into account, as well as the turning direction of the vehicle and the initial yaw rate and the maximum change in steering angle combined with the vehicle length. Subsequently, Section IV-B analyzes and determines all possible types of trajectories that directly follow from our equations. Each trajectory is characterized formally and demonstrated by example trajectories. At last, Section V discusses the applicability and further research regarding the models defined in this paper.

II. Definitions for the Two Models and Their Differences

In this section, we describe all symbols of the two models and common assumptions.

Physical model values are denoted as regular latin letters, while angles are denoted as greek letters. Symbols used in this paper are summarized in the following Table I.

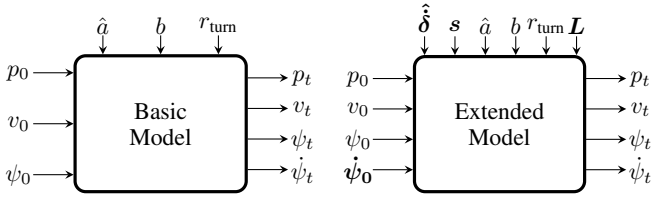
TABLE I. SYMBOLS AND NOTATION USED IN THIS PAPER.

Symbol	*Description*	*Unit*
X_i	Model state at time i	–
p	Position $\in \mathbb{R}^2$	m
X_{stop}	Stop state, $v = 0$	–
ψ	Yaw Angle	rad
$\dot{\psi}$	Yaw Rate	rad/s
s	Direction of steering as sign ± 1	–
$\dot{\hat{\delta}}$	Maximal steering angle change	rad/s
δ	Steering angle	rad
b	Braking Factor	–
$\hat{a}$	Maximum admissible acceleration	m/s^2
r_{turn}	Minimum turning radius	m
L	Length of a vehicle	m
$\mathcal{I}_\bullet$	Interval of admissible values for $\bullet$	–
$\bullet_{\min}, \bullet_{\max}$	Extreme values of $\mathcal{I}_\bullet$	–
$f_{\text{T}}(t), f_{\text{F}}(t), f_{\text{R}}(t)$	Part of $f(t)$ limited by **T**urning, **F**riction, **R**adius	–
$\widetilde{F}(t)$	Integral without constant: $\widetilde{F}(t) = \int f(t)dt - C_{\text{F}}$	–
$\overline{f}(t)$	Linearized version of $f(t)$	–
$f_x = f(t_x)$	Abbreviating function values of f at times t_x	–
$\mathcal{S}(t), \mathcal{C}(t)$	Sine and Cosine Fresnel Function[17]	–

Both models assume an ordered priority, on which the trajectory calculations are based upon:

1) The vehicle needs to be brought into a state with velocity $v = 0$ quickly.
2) Steering dynamics may be used to avoid obstacles or to reduce unavoidable impact.

In Figure 3, the main difference of the two models can be seen. The Extended Model takes the initial yaw rate $\dot{\psi}_0$ into consideration, which leads to three additional model parameters compared to the Basic Model, namely the maximum steering angle change rate $\dot{\hat{\delta}}$, the direction of steering as sign s and the vehicle length L.



(a) Basic Model without initial yaw rate and steering dynamics.

(b) Extension with one additional input and three additional parameters.

Figure 3. A comparison of both model's parameters (top) and state inputs (left). Note that model outputs (right) are the same in both models.

III. BASIC MODEL

This section describes the Basic Model, in which no initial yaw rate $\dot{\psi}_0$ is incorporated, however a maximal steering angle change rate $\dot{\hat{\delta}}$ of ∞ is assumed, which is sufficient for reachability estimation. A black box view of the model is shown in Figure 3a.

A. Equations of the Basic Model

The Basic Model is based on the so called *Friction Circle*, e.g., described by Pacejka [18]. As the modeled vehicle is braking in order to come to a full stop quickly, it will always be located near the boundary of this circle, either due to braking only, or by braking and turning in combination, as shown in Figure 4. The circle defines controllability when Equation (1) holds, where $\vec{a}_{\text{lon}}$ is the longitudinal acceleration component and $\vec{a}_{\text{cen}}$ the centripetal component, respectively.

$$\hat{a} \geq \|\vec{a}_{\text{res}}\| = \|\vec{a}_{\text{lon}} + \vec{a}_{\text{cen}}\| \tag{1}$$

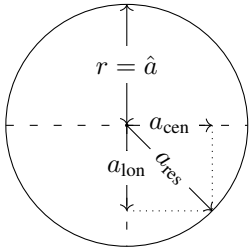


Figure 4. Friction Circle in the a_x, a_y-plane. Radius r is equal to the maximally applicable acceleration $\hat{a}$ between vehicle and road surface.

As the acceleration a_{res} results from a combination of braking and steering, the ratio $a_{\text{lon}}/\hat{a}$ causes different trajectories. We define this ratio by the factor b, as declared in Equation (2), further on called *Braking Factor*. We call b Braking Factor, as it describes the percentage of $\hat{a}$ that is applied for braking rather than turning. A b value of -0.5 means that 50% of the applicable acceleration is applied for braking. Note that $\hat{a}$ is positive, but when braking a_{lon} is negative, hence we choose $b \in [-1, 0]$.

$$b := \frac{a_{\text{lon}}}{\hat{a}} \tag{2}$$

The Basic Model provides a formal description of vehicle position $p(t) = [x, y]^T$ which is generally defined by the following integral:

$$p(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \int v(t) \begin{bmatrix} \cos(\psi(t)) \\ \sin(\psi(t)) \end{bmatrix} dt \tag{3}$$

where

$$v(t) = a_{\text{lon}} t + v_0 \tag{4}$$

A definition of $\psi(t)$ can be found by integrating the yaw rate $\dot{\psi}(t)$ over time t. In this Basic Model, the yaw rate is constrained by two different limits. The first limit is the Friction Circle, which does not allow higher yaw rates due to an otherwise resulting instable trajectory. This yaw rate is called $\dot{\psi}_{\text{F}}$. The second limit for the yaw rate is caused by the minimum turning radius r_{turn}. We call it $\dot{\psi}_{\text{R}}$. We thus define the yaw rate stepwise as given in Equation (5).

$$\dot{\psi}(t) = \begin{cases} \dot{\psi}_{\text{F}}(t) = \frac{\hat{a}\sqrt{1-b^2}}{v(t)}, & 0 \leq t \leq t_{\text{FR}} \\ \dot{\psi}_{\text{R}}(t) = \frac{v(t)}{r_{\text{turn}}}, & t_{\text{FR}} < t \leq t_{\text{stop}} \end{cases}, \forall t \geq 0 \tag{5}$$

where t_{FR} is defined as the time of intersection between $\dot{\psi}_{\text{F}}(t)$ and $\dot{\psi}_{\text{R}}(t)$ as can be seen in Equation (6) and t_{stop} is the time when $v = 0$ (see Equation (7)).

$$t_{\text{FR}} = \max\left(a_{\text{lon}}^{-1}\left(\sqrt{r_{\text{turn}}\hat{a}\sqrt{1-b^2}} - v_0\right), 0\right) \tag{6}$$

$$t_{\text{stop}} = -v_0 a_{\text{lon}}^{-1} \tag{7}$$

The yaw angle over time is then simply the time integral over $\dot{\psi}(t)$, as shown in Equation (8).

$$\psi(t) = \int \dot{\psi}(t)dt = \begin{cases} \psi_{\mathrm{F}}(t), & 0 \leq t \leq t_{\mathrm{FR}} \\ \psi_{\mathrm{R}}(t), & t_{\mathrm{FR}} < t \leq t_{\mathrm{stop}} \end{cases}, \forall t \geq 0 \quad (8)$$

where

$$\psi_{\mathrm{F}}(t) = z(\ln(v(t)) - \ln(v_0)) + \psi_0 \quad (9)$$

$$\psi_{\mathrm{R}}(t) = \left(\frac{1}{2}a_{\mathrm{lon}}t^2 + v_0 t\right) r_{\mathrm{turn}}^{-1} + C_{\psi,\mathrm{R}} \quad (10)$$

$$z = b^{-1}\sqrt{1-b^2} \quad (11)$$

$$C_{\psi,\mathrm{R}} = \psi_{\mathrm{F}}(t_{\mathrm{FR}}) - \left(\frac{1}{2}a_{\mathrm{lon}}t^2 - v_0 t\right) r_{\mathrm{turn}}^{-1} \quad (12)$$

Here, $C_{\psi,\mathrm{R}}$ is the constant of integration. With these equations, a solution for the positional integrals $x(t)$ and $y(t)$ can be found as shown in Equations (15) and (20). The x-Position can be calculated by Equations (13) and (14):

$$x_{\mathrm{F}}(t) = \frac{v(t)^2\left(z\sin(\psi(t)) + 2\cos(\psi(t))\right)}{a_{\mathrm{lon}}\left(z^2+4\right)} + C_{\mathrm{x,F}} \quad (13)$$

$$x_{\mathrm{R}}(t) = r_{\mathrm{turn}}\sin(\psi(t)) + C_{\mathrm{x,R}} \quad (14)$$

These equations describe position over time $x(t)$ and $y(t)$. See stepwise Equation (15) for $x(t)$.

$$x(t) = \begin{cases} x_{\mathrm{F}}(t), & 0 \leq t \leq t_{\mathrm{FR}} \\ x_{\mathrm{R}}(t), & t_{\mathrm{FR}} < t \leq t_{\mathrm{stop}} \end{cases} \quad (15)$$

The constant $C_{\mathrm{x,F}}$ is bound by the conditions $x(0) = x_0$, which means the vehicle must be at the starting position at time t_0. The constant for x_{F}, $C_{\mathrm{x,F}}$, is bound to hold the condition $x_{\mathrm{R}}(t_{\mathrm{FR}}) = x_{\mathrm{F}}(t_{\mathrm{FR}})$, which means that $x_{\mathrm{F}}(t)$ must seamlessly – e.g., in value and gradient – be continued by $x_{\mathrm{R}}(t)$ at t_{FR}. The result for both constants is described by Equations (16) and (17).

$$C_{\mathrm{x,F}} = x_0 - \frac{v_0^2\left(z\sin(\psi_0) + 2\cos(\psi_0)\right)}{a_{\mathrm{lon}}(z^2+4)} \quad (16)$$

$$C_{\mathrm{x,R}} = x(t_{\mathrm{FR}}) - r_{\mathrm{turn}}\sin(\psi(t_{\mathrm{FR}})) \quad (17)$$

The general description for $y(t)$ is shown below in Equation (20), while the step wise segments of all y-positions are shown in Equations (18) and (19):

$$y_{\mathrm{F}}(t) = -\frac{v(t)^2\left(z\cos(\psi(t)) - 2\sin(\psi(t))\right)}{a_{\mathrm{lon}}\left(z^2+4\right)} + C_{\mathrm{y,F}} \quad (18)$$

$$y_{\mathrm{R}}(t) = -r_{\mathrm{turn}}\cos(\psi(t)) + C_{\mathrm{y,R}} \quad (19)$$

$$y(t) = \begin{cases} y_{\mathrm{F}}(t), & 0 \leq t \leq t_{\mathrm{FR}} \\ y_{\mathrm{R}}(t), & t_{\mathrm{FR}} < t \leq t_{\mathrm{stop}} \end{cases} \quad (20)$$

The result for both constants to ensure seamlessness is described by Equations (21) and (22).

$$C_{\mathrm{y,F}} = y_0 - \frac{v_0^2\left(z\cos(\psi_0) - 2\sin(\psi_0)\right)}{a_{\mathrm{lon}}(z^2+4)} \quad (21)$$

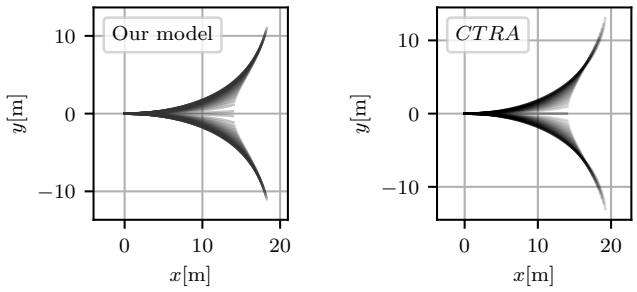
$$C_{\mathrm{y,R}} = y(t_{\mathrm{FR}}) + r_{\mathrm{turn}}\cos(\psi(t_{\mathrm{FR}})) \quad (22)$$

The trajectory of a braking and turning vehicle is described as $p(t)$, by the compound x- and y-position in Cartesian coordinates over time t.

Before introducing the model extension, we discuss the Basic Model and compare it with the CTRA Model in Section III-B, in order to understand the implications of the approximations and model parameters first.

B. Comparison of Our Basic Model Against CTRA Model

To evaluate the Basic Model's performance with respect to calculation time and to show its correctness, we compare it to a CTRA-Model [19] (*Constant Turn Rate and Acceleration*) in a simulation. The CTRA simulation iteratively moves a vehicle, such that our condition in (1) is fulfilled, and the assumptions introduced in Section II hold. The simulation therefore calculates effectively the same maneuvers as our Basic Model, but in a very different way. We choose the CTRA-Model, as it is well known, allows the vehicle to follow a spiral shape and has the same state space representation as our model. The turn rate and acceleration is assumed to be constant within one of many consecutive time steps Δt.



(a) Our basic braking model. (b) CTRA-Model, $\Delta t = 0.0075s$.

Figure 5. Comparison of our Basic Model to the CTRA-Model for 40 vehicle trajectories with linearly sampled b values and equal start state.

The result in Figure 5 shows that our model matches the shape of the CTRA-Model well, without introducing linearization errors as the CTRA-Model does. Both results from Figure 5 show a very similar structure. The starting conditions for both tests are $v_0 = 16.67^{m}/_{s}, \hat{a} = 10^{m}/_{s^2}, r_{\mathrm{turn}} = 12.5m, \psi_0 = 0\ rad$. Note that the CTRA-Model (Figure 5b) has slightly longer trajectories, especially in the outer arms of the structure. This is caused by the CTRA-Model's assumption of a constant turn rate $\dot{\psi}$, which is not correct in this kind of non-linear maneuver. In our model (Figure 5a), the only assumption is that of a constant acceleration, as introduced in Section II.

The main advantage of our model is the fact that we can directly compute certain vehicle positions straight from the formulas derived in Section III-A such that time intensive calculations are not necessary. A comparison of computation times t_{calc} in seconds, and their deviation $\sigma_{t_{\mathrm{calc}}}$ over 10 runs is shown in Table II. In the first test, only the stop states where

computed of 1000 different b values. In the second test, a whole pearl chain of positions from start to stop was computed, with 250 points per b value.

TABLE II. COMPARISON TO THE CTRA MODEL.

Calculate 1000 possible **stop states**, $\Delta t = 0.01112s$						
v_0	5 m/s		10 m/s		20 m/s	
	Mean t_{calc}[s]	$\sigma_{t_{\text{calc}}}$	Mean t_{calc}	$\sigma_{t_{\text{calc}}}$	Mean t_{calc}	$\sigma_{t_{\text{calc}}}$
CTRA	1.0715	0.0137	2.1975	0.0052	4.9310	0.1073
Our model	0.2059	0.0053	0.2078	0.0017	0.2144	0.0075
Calculate 1000 **trajectories**, 250 samples per trajectory, $\Delta t = 0.01112s$						
v_0	5 m/s		10 m/s		20 m/s	
	Mean t_{calc}	$\sigma_{t_{\text{calc}}}$	Mean t_{calc}	$\sigma_{t_{\text{calc}}}$	Mean t_{calc}	$\sigma_{t_{\text{calc}}}$
CTRA	1.0870	0.0207	2.2335	0.0096	4.9761	0.0814
Our model	0.2310	0.0017	0.2326	0.0021	0.2320	0.0011

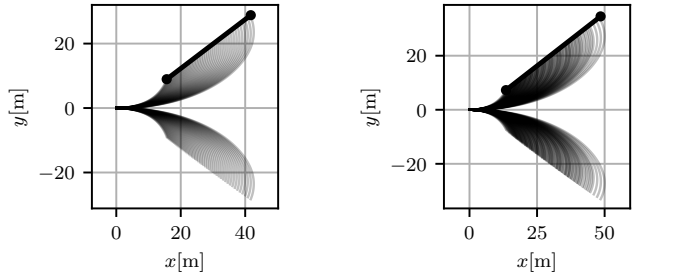
The table shows that our model is up to 20 times faster in terms of computing time than the CTRA-Model, especially for high initial velocities v_0. This is caused by the fact that CTRA must iteratively compute time steps until the stop position is found, whereas our model can directly compute the stop state.

C. Discussion of Model Uncertainties

In this section, we discuss the effect of individual uncertainties in the model parameters $r_{\text{turn}}, \hat{a}$ and the initial vehicle state $X_0 = [x_0, y_0, v_0, \psi_0]^T$. We model the uncertainties as intervals $\mathcal{I}_\Theta$,$\mathcal{I}_{X_0}$ that contain all possible values. As the parameters are also contained in the Extended Model in Section IV, this discussion is valid for both models.

1) Highest possible deceleration $\hat{a}$: The highest possible deceleration heavily depends on the road and tire conditions, which are often uncertain. The interval $\mathcal{I}_{\hat{a}}$ therefore covers the most slippery and most rough road condition possible. Calculating different stop states X_{stop} with different values for $\hat{a}$ reveals an almost linear behavior within expectable values of $\hat{a} \in \mathcal{I}_{\hat{a}}$.

The resulting shape of 50 different $\hat{a} \in \mathcal{I}_{\hat{a}}$ can be seen in Figure 6a, where lower values of $\hat{a}$ lead to a farther vehicle trajectory with an almost linear behavior. A line segment shows the extending effect of the parameter uncertainties on the top half.



(a) Resulting trajectories at interval $\mathcal{I}_{\hat{a}} = [4, 12]\ \frac{m}{s^2}$.

(b) Resulting trajectories at intervals $\mathcal{I}_{\hat{a}} = [4, 12]\ \frac{m}{s^2}, \mathcal{I}_{v_0} = [15.3, 18.1]\ \frac{m}{s}$

Figure 6. Two sets of trajectories with a b value of -0.6. Left, only considering $\mathcal{I}_{\hat{a}}$. Right, considering $\mathcal{I}_{\hat{a}}$ and $\mathcal{I}_{v_0}$.

2) Smallest possible turning radius r_{turn}: The smallest possible turning radius r_{turn} is a vehicle inherent parameter which influences the trajectory after t_{FR} and also defines the value of t_{FR} itself. Although there are certain legal requirements for r_{turn} depending on vehicle class, the exact value is uncertain, especially when considering other traffic participants.

Any $r_{\text{turn}} \in \mathcal{I}_{r_{\text{turn}}}$ causes a different stopping position. Unfortunately, neither the lowest nor the highest r_{turn} necessarily leads to the outmost stopping position. By observing the stopping positions depending on r_{turn}, one can see that the shape of all stopping positions with different $r_{\text{turn}} \in \mathcal{I}_{r_{\text{turn}}}$ forms a spiral with a rising radius. Let A be the stopping position of the lowest r_{turn}, $A = X_{\text{stop}}|_{r_{\text{turn,min}}}$, and $B = X_{\text{stop}}|_{r_{\text{turn,max}}}$. The circle with radius $r = dist(A, B)$ at center A then includes all points of the spiral, which means all stopping positions can be overestimated by such a circle. By describing this distance as function $d = f(\hat{a}, v_0)$, it can be shown that the maximum distance is at $d_{\max} = f(\hat{a}_{\min}, v_{0,\max})$. Figure 7 shows an example of such a circle.

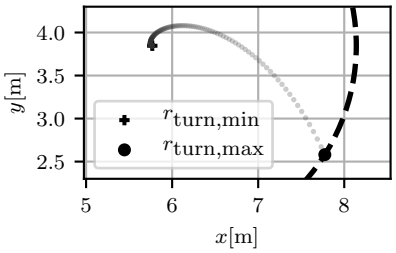


Figure 7. Effect of $\mathcal{I}_{r_{\text{turn}}}$ on X_{stop}. The figure shows how a circle can surround all stopping positions caused by different $r_{\text{turn}} \in \mathcal{I}_{r_{\text{turn}}} = [1\text{e}{-7}, 13]m$.

In order to show the spiral effect in Figure 7, we assumed $\mathcal{I}_{r_{\text{turn}}} = [1\text{e}{-7}, 13]m$ and $v_0 = 10^m/_s$, which results in a circle radius of $\approx 2.4m$. For a more realistic scenario of $\mathcal{I}_{r_{\text{turn}}} = [7, 13]m$ and $v_0 = 10^m/_s$, the radius of the circle is $\approx 1.3m$.

3) Initial velocity v_0: The uncertainty in the initial velocity $\mathcal{I}_{v_0}$ determines the stopping distance similarly to $\mathcal{I}_{\hat{a}}$, as it stretches the possibly reachable positions farther from the start. This means the closest reachable position is defined by $v_{0,\min}$ and $\hat{a}_{\max}$, which stands for a very rough road-to-tire surface. In contrast, the farthest reachable stopping position is defined by the highest velocity $v_{0,\max}$ on the most slippery road $\hat{a}_{\min}$ possible. An example of the resulting shape is shown in Figure 6b.

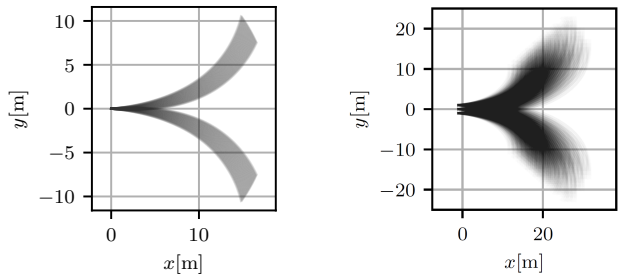
4) Initial position: The initial position of the vehicle will always be uncertain, as no perfect localization is possible. The effect of an uncertain starting position (x_0, y_0) is however not complex, as a different starting position of $\Delta x, \Delta y$ simply causes a translation of the complete reachable area of $\Delta x, \Delta y$.

5) Initial yaw angle: The initial yaw angle rotates the complete reachable area around the starting position of the vehicle. Figure 8a shows an example of this effect, where $\mathcal{I}_{\psi_0} = [-\pi/32, \pi/32]$.

6) Combination of all uncertainties: So far, we discussed the uncertainty of parameters separately. To describe and overestimate all system states that can potentially be reached under all uncertainties is not in the scope of this paper. In order to do so, a formal reachability analysis must be performed, compare for example [6][8][12][20].

By sampling all parameters from $\mathcal{I}$ and calculating all combinations, we can estimate the reachable area non formally by the union of the resulting shapes.

In Figure 8b we show such a result, where $\mathcal{I}_{\hat{a}} = [7, 11]$, $\mathcal{I}_{r_{\text{turn}}} = [7, 13]$, $\mathcal{I}_{v_0} = [15.3, 18.1]$, $\mathcal{I}_{\psi_0} = [-\pi/32, \pi/32]$, $\mathcal{I}_{x_0} = \mathcal{I}_{y_0} = [-1, 1]$. We sample 3 parameters of each interval. The above section show how the Basic Model of this paper can be used to estimate the Braking Area. To take into account the physics of a limited change rate, we extend the Basic Model by introducing the additional parameters $s, L, \hat{\dot{\delta}}$ and the new model input of $\dot{\psi}_0$ as shown in Figure 3b.



(a) Trajectories at interval $\mathcal{I}_{\psi_0} = [-\pi/32, \pi/32]$. Other parameters, $b = -0.6$, $\hat{a} = 4m/s^2$, $v_0 = 15.3m/s$, $r_{\text{turn}} = 12.5m$.

(b) Trajectories at interval $\mathcal{I}_{\hat{a}} = [7, 11]m/s^2$, $\mathcal{I}_{r_{\text{turn}}} = [7, 13]m$, $\mathcal{I}_{v_0} = [15.3, 18.1]m/s$, $\mathcal{I}_{\psi_0} = [-\pi/32, \pi/32]$rad, $\mathcal{I}_{x_0} = \mathcal{I}_{y_0} = [-1, 1]m$.

Figure 8. The effect of uncertain parameters. Left, only $\mathcal{I}_{\psi_0}$ is considered. Right, all parameters are assumed uncertain.

IV. Extended Model

The missing yaw rate at $t = 0$ permits estimating the reachable area of the vehicle, however, leads to inaccurate trajectory results in situations in which a vehicle has an initial yaw rate $\dot{\psi}_0$ not close or equal to 0, especially when steering is slow.

For the Extended Model, the general positional integrals remain the same as in Equation (3) introduced in the Basic Model. The difference lies in the calculation of yaw angle $\psi(t)$. Instead of directly applying the highest yaw rate permitted by the Friction Circle $\dot{\psi}_{\text{F}}$, we instead start from the current yaw rate at $t = 0$, called $\dot{\psi}_0$. From this point forward, the yaw rate is computed by steering, until the vehicle either reaches its stopping position, or until another limitation is reached.

A. Equations of the Extended Model

This yields three different descriptions of the current yaw rate. In general, the yaw rate is defined in Equation (23) as:

$$\dot{\psi}(t) = \frac{v(t)}{r(t)} = v(t)\kappa(t) \tag{23}$$

where $r(t)$ is the radius of the trajectory and $\kappa(t)$ the respective curvature. An advantage of curvature over radius is that a straight curve has a radius of ∞, but a curvature of 0, which is much easier for computations.

Hence, the difference between the trajectory segments lies in the different descriptions of curvatures, which can be computed trivially.

Figure 9 shows an example of the different curvatures.

In the following, all different curvature types $\kappa_{\text{T}}(t)$, $\kappa_{\text{F}}(t)$ and $\kappa_{\text{R}}(t)$ will be described in detail.

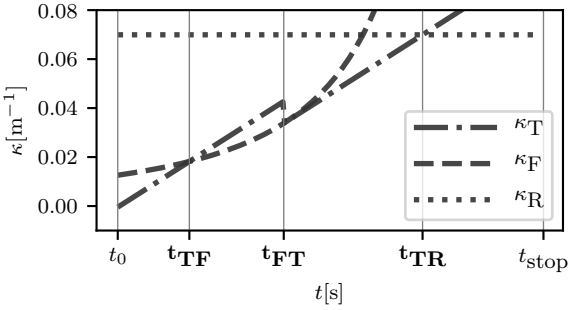


Figure 9. The 3 curvatures $\kappa_{\text{T}}, \kappa_{\text{F}}, \kappa_{\text{R}}$. The minimum of all curves defines the biggest possible curvature in each point in time between $t = 0$ and $t = t_{\text{stop}}$.

Curvature Defined by Change of Steering Angle: The curvature of road vehicles is determined by the steering angle δ, which is set by the steering wheel and is defined by the angle of vehicle longitudinal axis and the direction the wheels are pointing. Because of mechanical and physical limits, the steering wheel cannot be moved arbitrarily fast. To incorporate this property in our Extended Model, we introduce the new parameter $\hat{\dot{\delta}}$, which describes the maximum steering angle change rate. The resulting curvature definition based on the steering angle δ and maximum change rate $\hat{\dot{\delta}}$ is shown in Equation (24):

$$\kappa_{\text{T,l}}(t) = \frac{\delta(t)}{L} = \frac{s\hat{\dot{\delta}}t + \delta_0}{L} \tag{24}$$

where the steering angle at $t = 0$, δ_0, is defined in Equation (25) as:

$$\delta_0 = \frac{\dot{\psi}_0 L}{v_0} \tag{25}$$

However, the description of $\kappa_{\text{T,l}}(t)$ is not sufficient to describe $\kappa_{\text{T}}(t)$ completely. The reason can be seen in Figure 9 in the appearance of κ_{T}, specifically in the jump that the curve performs at t_{FT}. The maximum change rate of steering angle, $\hat{\dot{\delta}}$, not only becomes effective from $t = 0$ onwards, but also at any other times. However, the only situation of such kind, in the model, is when the Friction Circle allows for change in curvature which is greater than $\hat{\dot{\delta}}$. In that case, it must be limited by a differently defined segment of $\kappa_{\text{T}}(t)$, called $\kappa_{\text{T,r}}(t)$. This segment $\kappa_{\text{T,r}}(t)$ is defined as shown in Equation (26), which means that the original curve $\kappa_{\text{T,l}}$ is shifted by an offset o_{T} along the y-axis.

$$\kappa_{\text{T,r}} = \kappa_{\text{T,l}} + o_{\text{T}} \tag{26}$$

The offset o_{T} must ensure that the gradient of $\kappa_{\text{F}}(t)$ does not exceed the gradient of $\kappa_{\text{T}}(t)$. It is defined in Equation (27) as:

$$o_{\text{T}} = \kappa_{\text{F}}(t_{\text{FT}}) - \kappa_{\text{T,l}}(t_{\text{FT}}) \tag{27}$$

where

$$t_{\text{FT}} = -{a_{\text{lon}}}^{-1}\left(\frac{\sqrt[3]{2\hat{a}^2 bL\sqrt{1-b^2}}}{\sqrt[3]{\hat{\dot{\delta}}}} + v_0\right) \tag{28}$$

With this, a complete description of $\kappa_{\text{T}}(t)$ can be formulated as Equation (29):

$$\kappa_{\text{T}}(t) = \min\left(\kappa_{\text{T,l}}(t), \kappa_{\text{T,lr}}(t)\right) \tag{29}$$

where

$$\kappa_{\mathrm{T,lr}}(t) = \begin{cases} \kappa_{\mathrm{T,l}}(t), & 0 \leq t \leq t_{\mathrm{FT}} \\ \kappa_{\mathrm{T,r}}(t), & t_{\mathrm{FT}} < t < t_{\mathrm{stop}} \end{cases} \tag{30}$$

In words, the above Equation (29) means that if $\kappa_{\mathrm{T,l}}$ does not intersect κ_{F}, there is no second segment of κ_{T}. However, if it does intersect, the two-segmented description from Equation (30) shall be used.

The yaw rate based on the initial yaw rate $\dot{\psi}_0$ is hence defined as shown in Equation (31).

$$\dot{\psi}_{\mathrm{T}}(t) = v(t)\kappa_{\mathrm{T}} \tag{31}$$

$\dot{\psi}_{\mathrm{T}}(t)$ describes the yaw rate of a vehicle that is constantly turning as quickly as the mechanical system allows it.

As can be seen in Equation (31), the yaw rate described by constantly changing the steering angle results in a second order polynomial.

If this polynomial was used to further find a description of position, a computationally too expensive intermediate result emerges. The reason for that can be found in $\psi_{\mathrm{T}}(t)$, which naturally is a third order polynomial, see Equation (32).

$$\psi_{\mathrm{T}}(t) = \int \dot{\psi}_{\mathrm{T}}(t)dt = At^3 + Bt^2 + Ct + D \tag{32}$$

where

$$\begin{aligned} A &= \frac{\hat{a}b\hat{\dot{\delta}}s}{3L} \\ B &= \frac{3\hat{a}b\delta_0 + 3\hat{\dot{\delta}}sv_0}{6L} \\ C &= \frac{\delta_0 v_0}{L} \\ D &= \text{const.} \end{aligned}$$

As the position is defined as Equation (33), it means that an integral in the form of $\int t\cos(t^3)dt$ must be solved.

$$p_{\mathrm{T}}(t) = \begin{bmatrix} x_{\mathrm{T}}(t) \\ y_{\mathrm{T}}(t) \end{bmatrix} = \int v(t) \begin{bmatrix} \cos(\psi_{\mathrm{T}}(t)) \\ \sin(\psi_{\mathrm{T}}(t)) \end{bmatrix} dt \tag{33}$$

The integral of Equation (33) leads to a large combination of Incomplete Gamma Functions as described, e.g., by Paris [21, 8]. Though several asymptotic approximations exist, see [21, 8.25], the computational complexity is expected to overrule the resulting gain in accuracy. Hence, we introduce an approximation earlier in the position calculation. Instead of applying the parabolic shape of $\dot{\psi}_{\mathrm{T}}(t)$, we approximate the relevant parabola segment by straight line segments. See the first bold line segment in Figure 10 for an example.

Note that the approximation always under-estimates the actually possible yaw rate and is hence feasible. The formal definition of any approximative line from t_1 to t_2 is shown in Equation (34).

$$\overline{\dot{\psi}_{\mathrm{T}}}(t|t_1, t_2) = \frac{\dot{\psi}(t_2) - \dot{\psi}(t_1)}{t_2 - t_1}\, t + \frac{t_2\dot{\psi}(t_1) - t_1\dot{\psi}(t_2)}{t_2 - t_1} \tag{34}$$

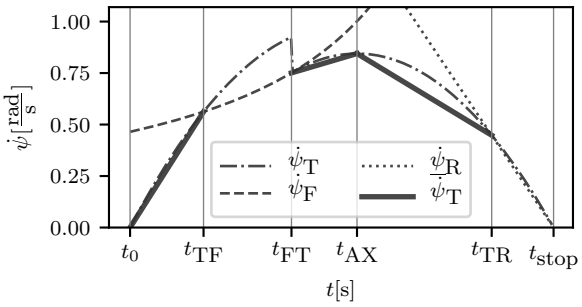


Figure 10. An example of linearizing $\dot{\psi}_{\mathrm{T}}(t)$. Two parabola segments are approximated, the first from 0 to t_{TF}. The second one from t_{FT} to t_{TR}.

where $\dot{\psi}(t_{1,2}) \in \dot{\psi}_{\mathrm{T}}$. Note that even though the chosen yaw rates at $t_{1,2}$ must be $\in \dot{\psi}_{\mathrm{T}}$, they can also be computed with another yaw rate description, as $t_{1,2}$ are intersection times of yaw rates, and hence the yaw rates of at least one other segment of $\dot{\psi}(t)$ are identical at these times.

The second line segment from t_{FT} to t_{TR} in Figure 10 shows a special case that must be considered when linearizing $\dot{\psi}_{\mathrm{T}}$. The parabola segment spans across the parabola's Apex at t_{AX}. See Figure 11 for a more detailed example. If unconsidered, this case could cause large errors.

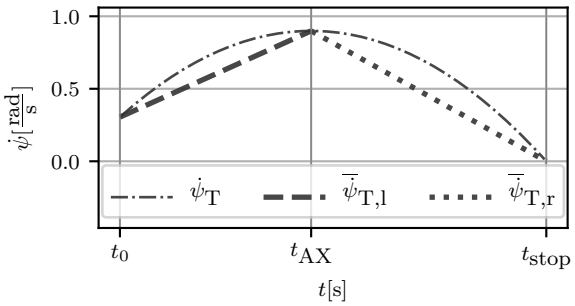


Figure 11. Linearization of $\dot{\psi}_{\mathrm{T}}(t)$ by two line segments $\overline{\dot{\psi}_{\mathrm{T,l}}}$ and $\overline{\dot{\psi}_{\mathrm{T,r}}}$ in the special case that the line segment spans across the Apex Point at time t_{AX}.

The Apex Point is defined in Equation (35), where o_{T} is the offset of the second segment of $\kappa_{\mathrm{T,lr}}$, as introduced in Equation (30).

$$t_{\mathrm{AX}} = -\frac{\hat{a}b\left(\delta_0 + Lo_{\mathrm{T}}\right) + \hat{\dot{\delta}}v_0}{2\hat{a}b\hat{\dot{\delta}}} \tag{35}$$

With these descriptions of approximated yaw rates $\overline{\dot{\psi}_{\mathrm{T}}}$, a resulting description for the yaw angle ψ_{T} can be formulated, now as a second order polynomial as shown in Equation (36).

$$\overline{\psi_{\mathrm{T}}}(t|t_1, t_2) = \frac{\Delta\dot{\psi}}{2\Delta t}t^2 + \frac{\dot{\psi}(t_1)t_2 - \dot{\psi}(t_1)t_1}{\Delta t}t + \psi(t_1) \tag{36}$$

where

$$\begin{aligned} \Delta t &= t_2 - t_1 \\ \Delta\dot{\psi} &= \dot{\psi}(t_2) - \dot{\psi}(t_1) \end{aligned}$$

The compound position Equation (37) is then

$$\begin{aligned}&\overline{p_{\mathrm{T}}}(t|t_1,t_2) = \\ &= \begin{bmatrix} x_{\mathrm{T}}(t) \\ y_{\mathrm{T}}(t) \end{bmatrix} = \int v(t) \begin{bmatrix} \cos(\overline{\psi_{\mathrm{T}}}(t)) \\ \sin(\overline{\psi_{\mathrm{T}}}(t)) \end{bmatrix} dt \\ &= \begin{bmatrix} +\sigma_0 \sin(\xi_0) + \sigma_1(\sin(\xi_1)\mathbb{S} + \cos(\xi_1)\mathbb{C}) \\ -\sigma_0 \cos(\xi_0) + \sigma_1(\cos(\xi_1)\mathbb{S} - \sin(\xi_1)\mathbb{C}) \end{bmatrix} \sigma_2 + C_{\mathrm{p,T}}\end{aligned} \tag{37}$$

where

$$\sigma_1 = \sqrt{\pi \Delta t} \left(\Delta \dot{\psi} v_0 + a_{\mathrm{lon}} \left(\dot{\psi}_2 t_1 - \dot{\psi}_1 t_2 \right) \right)$$

$$\sigma_0 = \sqrt{\Delta \dot{\psi}} a_{\mathrm{lon}} \Delta t \quad \sigma_2 = \Delta \dot{\psi}^{-\frac{3}{2}}$$

$$\xi_0 = \frac{\frac{1}{2}\Delta \dot{\psi} t^2 + \left(\dot{\psi}_1 t_2 - \dot{\psi}_2 t_1 \right) t + \psi_1 t_2 - \dot{\psi}_1 t_1}{\Delta t}$$

$$\xi_1 = \frac{\dot{\psi}_1^2 t_2^2 + \dot{\psi}_2^2 t_1^2 - 2 t_2 \left(\Delta \dot{\psi} \psi_1 + \dot{\psi}_1 \dot{\psi}_2 t_1 \right) + 2 \Delta \dot{\psi} \psi_1 t_1}{2 \Delta \dot{\psi} \Delta t}$$

$$\mathbb{C} = \mathcal{C}(\sigma_3) \quad \mathbb{S} = \mathcal{S}(\sigma_3) \quad \sigma_3 = \frac{\Delta \dot{\psi} t + \dot{\psi}_1 t_2 - \dot{\psi}_2 t_1}{\sqrt{\pi \Delta \dot{\psi} \Delta t}}$$

The constant $C_{\mathrm{p,T}}$ must be calculated specific to the type of trajectory, see Section IV-B for more. Once turning reaches the boundary of the Friction Circle, another set of equations limits the yaw rate and hence defines the yaw angle and the position.

Curvature Defined by Friction Circle: The equations are very similar to the Basic Model, as the Physics are the same. The curvature defined by the Friction Circle is more complex than the linear relation of $\kappa_{\mathrm{T}}(t)$, as can be seen in Equation (38).

$$\kappa_{\mathrm{T}} = s \frac{\hat{a}\sqrt{1-b^2}}{v(t)^2} \tag{38}$$

Note that in addition to the Basic Model, the extended version includes the Steering Direction s. The yaw angle is less complex, as one of the $v(t)$ terms gets removed, see Equation (39).

$$\dot{\psi}_{\mathrm{F}}(t) = s \frac{\hat{a}\sqrt{1-b^2}}{v(t)} \tag{39}$$

Respectively, the yaw angle $\psi_{\mathrm{F}}(t)$ in the Extended Model is shown in Equation (40).

$$\psi_{\mathrm{F}}(t) = sz(\ln(v(t)) - \ln(v_0)) + C_{\dot{\psi},\mathrm{F}} \tag{40}$$

The position calculation is shown in Equation (41):

$$\begin{aligned}p_{\mathrm{F}}(t|t_1) &= \begin{bmatrix} x_{\mathrm{F}}(t) \\ y_{\mathrm{F}}(t) \end{bmatrix} = \int v(t) \begin{bmatrix} \cos(\psi_{\mathrm{F}}(t)) \\ \sin(\psi_{\mathrm{F}}(t)) \end{bmatrix} dt \\ &= \sigma_0 \left[sz \begin{bmatrix} +\sin(\psi_1) \\ -\cos(\psi_1) \end{bmatrix} + 2 \begin{bmatrix} +\cos(\psi_1) \\ -\sin(\psi_1) \end{bmatrix} \right] + C_{\mathrm{p,F}}\end{aligned} \tag{41}$$

where

$$\sigma_0 = v(t)^2 \left(a_{\mathrm{lon}}(z^2 s^2 + 4) \right)^{-1}$$

The constant of integration $C_{\mathrm{p,F}}$ must be adjusted and computed specific to the type of trajectory, see Section IV-B for more. In the Extended Model, the yaw rate can naturally also be limited by the smallest turning radius r_{turn}.

Curvature Defined by Turning Radius: The curvature defined by the turning radius is simply its reciprocal value times s, as shown in Equation (42).

$$\kappa_{\mathrm{R}} = s r_{\mathrm{turn}}^{-1} \tag{42}$$

Note that κ_{R} is independent of time, as it is only defined by properties of the vehicle. The resulting yaw rate is shown in Equation (43):

$$\dot{\psi}_{\mathrm{R}}(t) = s \frac{v(t)}{r_{\mathrm{turn}}} \tag{43}$$

The yaw angle is shown in Equation (44):

$$r_{\mathrm{turn}}^{-1} \left(\frac{1}{2} s a_{\mathrm{lon}} t^2 + s v_0 t \right) + C_{\psi,\mathrm{R}} \tag{44}$$

The resulting position defined by the turning radius in the Extended Model is shown in Equation (45):

$$\begin{aligned}p_{\mathrm{R}}(t|t_1) &= \begin{bmatrix} x_{\mathrm{R}}(t) \\ y_{\mathrm{R}}(t) \end{bmatrix} = \int v(t) \begin{bmatrix} \cos(\psi_{\mathrm{R}}(t)) \\ \sin(\psi_{\mathrm{R}}(t)) \end{bmatrix} dt \\ &= s r_{\mathrm{turn}} \begin{bmatrix} +\sin(\psi_1) \\ -\cos(\psi_1) \end{bmatrix} + C_{\mathrm{p,R}}\end{aligned} \tag{45}$$

In order to calculate trajectories, the different position descriptions must be interconnected to ensure seamlessness in terms of angle and position. This is done by appropriately choosing the constants of integration of all position and yaw angle equations. For that purpose, the intersection times of all curvatures play a major role. In the Basic Model, the intersection time is defined easily because there is only one. In contrast, the Extended Model has different intersection times, and their calculation is partially less trivial.

The Meaning of Curvature Intersections: The curvature of a resulting trajectory is limited by the three curves $\kappa_{\mathrm{T}}(t), \kappa_{\mathrm{F}}(t), \kappa_{\mathrm{R}}(t)$. For positive curvatures, as shown in Figure 9, the maximally curved trajectory considering all inputs is hence defined as shown in Equations (46) and (47).

$$\kappa(t) = \min(\kappa_{\mathrm{T}}(t), \kappa_{\mathrm{F}}(t), \kappa_{\mathrm{R}}) \forall t \in [0, t_{\mathrm{stop}}] \tag{46}$$

$$\dot{\psi}(t) = \min(\dot{\psi}_{\mathrm{T}}(t), \dot{\psi}_{\mathrm{F}}(t), \dot{\psi}_{\mathrm{R}}(t)) \forall t \in [0, t_{\mathrm{stop}}] \tag{47}$$

The yaw rate curves intersect at the same times as the curvatures, as long as $v(t) \neq 0$, because

$$\begin{aligned}\dot{\psi}_x &= \dot{\psi}_y && \Leftrightarrow \\ v(t)\kappa_x &= v(t)\kappa_y && \Leftrightarrow \\ \kappa_x &= \kappa_y && \forall v(t) \neq 0 \Leftrightarrow \forall t \neq t_{\mathrm{stop}}\end{aligned}$$

Therefore, the different sections of the Extended Model trajectory are each limited by the respective intersection times of the three curvatures. The intersection at t_{stop} is not relevant for that, as the model ends its calculations when velocity is 0. An overview of all yaw rates is given in Figure 12, which shows all different characteristics that describe the course of the yaw rates of a vehicle while braking and turning.

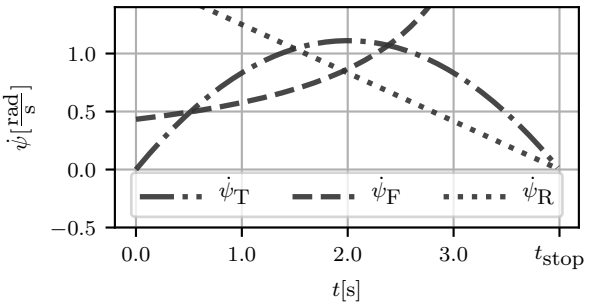


Figure 12. The 3 different yaw rate characteristics $\dot{\psi}_\text{T}$, $\dot{\psi}_\text{F}$, $\dot{\psi}_\text{R}$. The Extended Model assumes the minimum of all yaw rates at each point in time.

All yaw rate descriptions have in common that they describe the maximal yaw rate possible. That means that at all times t, the yaw rate is limited by all three descriptions.

As shown in the Basic Model, the yaw angle $\psi(t)$ is the integral over the yaw rate $\dot{\psi}(t)$ over time. In order to integrate $\dot{\psi}(t)$, a step wise definition is helpful, similar to Equation (5) of the Basic Model. We thus define $\dot{\psi}(t)$ as shown in Equation (48).

$$\dot{\psi}(t) = \begin{cases} \dot{\psi}_\text{T}(t), & 0 \leq t \leq t_\text{TF} \\ \dot{\psi}_\text{F}(t), & t_\text{TF} < t \leq t_\text{FR} \\ \dot{\psi}_\text{R}(t), & t_\text{FR} < t < t_\text{stop} \end{cases}, \forall t \geq 0 \tag{48}$$

where

$$\dot{\psi}_\text{T}(t) = v(t)(s\hat{\dot{\delta}}t + \delta_0)L^{-1} \tag{49}$$

$$\dot{\psi}_\text{F}(t) = s\hat{a}\sqrt{1-b^2}\ v^{-1}(t) \tag{50}$$

$$\dot{\psi}_\text{R}(t) = sv(t)r_\text{turn}^{-1} \tag{51}$$

where s is the direction of steering and only defines a positive or negative sign, positive meaning *turning left* and negative *turning right*.

From Figure 12 and Equation (48) follows that the model behavior is defined by the positions of intersections of the different yaw rate descriptions, and all limit the vehicle movement in their own way. Therefore, the definition of these times is very important. Simply put, these times are the times at which the different curvatures intersect.

The intersections of κ_T and $\kappa_\text{F} \in \mathbb{R}$ can be found by applying Cardano's formula [22] to the Cubic Equation (52):

$$At^3 + Bt^2 + Ct + D = 0 \tag{52}$$

where

$$A = s{a_\text{lon}}^2\hat{\dot{\delta}} \tag{53}$$

$$B = {a_\text{lon}}^2\delta_0 + 2sa_\text{lon}v_0\hat{\dot{\delta}} \tag{54}$$

$$C = 2a_\text{lon}v_0\delta_0 + s\hat{\dot{\delta}}{v_0}^2 \tag{55}$$

$$D = \delta_0{v_0}^2 - sL\hat{a}\sqrt{1-b^2} \tag{56}$$

The determinant is defined as Equation (57):

$$\Delta := \left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3 \tag{57}$$

where

$$p = \beta - \frac{\alpha^2}{3} \tag{58}$$

$$q = \frac{2\alpha^3}{27} - \frac{\alpha\beta}{3} + \gamma \tag{59}$$

$$u = \sqrt[3]{-\frac{q}{2} + \sqrt{\Delta}} \tag{60}$$

$$v = \sqrt[3]{-\frac{q}{2} - \sqrt{\Delta}} \tag{61}$$

with $\alpha = {}^A\!/_D$, $\beta = {}^B\!/_D$ and $\gamma = {}^C\!/_D$.

If $\Delta > 0$, there is only one intersection $\in \mathbb{R}$ at

$$t_\text{T,F,0} = u + v - \frac{B}{3A} \tag{62}$$

If $\Delta = 0$ and $p = 0$, there is a triple intersection $\in \mathbb{R}$ at

$$t_\text{T,F,0,1,2} = -\frac{B}{3A} \tag{63}$$

If $\Delta = 0$ but $p \neq 0$, there are three solutions $\in \mathbb{R}$ at

$$t_\text{T,F,0} = \frac{3q}{p} - \frac{B}{3A} \tag{64}$$

$$t_\text{T,F,1,2} = -\frac{3q}{2p} - \frac{B}{3A} \tag{65}$$

If $\Delta < 0$, there are also three solutions $\in \mathbb{R}$, which may be calculated by

$$t_\text{T,F,0} = \sqrt{-\frac{4}{3}p}\cos\left(\sigma_\text{T,F}\right) - \frac{B}{3A} \tag{66}$$

$$t_\text{T,F,1} = \sqrt{-\frac{4}{3}p}\cos\left(\sigma_\text{T,F} + \frac{\pi}{3}\right) - \frac{B}{3A} \tag{67}$$

$$t_\text{T,F,2} = \sqrt{-\frac{4}{3}p}\cos\left(\sigma_\text{T,F} - \frac{\pi}{3}\right) - \frac{B}{3A} \tag{68}$$

where

$$\sigma_\text{T,F} = \frac{1}{3}\arccos\left(-\frac{q}{2}\sqrt{-\frac{27}{p^3}}\right)$$

The other intersection times can be calculated generally as:

$$t_\text{F,R,0} = -\frac{v_0 + \sqrt{\hat{a}r_\text{turn}\sqrt{1-b^2}}}{a_\text{lon}} \tag{69}$$

$$t_\text{F,R,1} = -\frac{v_0 - \sqrt{\hat{a}r_\text{turn}\sqrt{1-b^2}}}{a_\text{lon}} \tag{70}$$

$$t_\text{T,R,0} = -\frac{v_0}{a_\text{lon}} = t_\text{stop} \tag{71}$$

$$t_\text{T,R,1} = -\frac{sL + \delta_0 r_\text{turn}}{\hat{s\dot{\delta}}r_\text{turn}} \tag{72}$$

Equations (62) to (72) describe all possible intersection times between the different yaw rate limits. Because the curvature defined by the Friction Circle κ_F is strictly increasing or decreasing in $[0, t_\text{stop}]$ (depending on s), the linear function $\kappa_\text{F}(t)$ intersects $\kappa_\text{T}(t)$ possibly at two times in the interval $[0, t_\text{stop}]$. Also, note that $\dot{\psi}_\text{T}$ intersects $\dot{\psi}_\text{R}$ also in t_stop, even though the curvatures are not equal, because the equality condition only holds for $v(t) \neq 0$, which however is the case at t_stop. See Figure 12 for an example.

B. Different Trajectory Types of the Extended Model

As shown above, the positions of yaw rate intersection define the course of a trajectory. An intersection time is called *relevant*, when at this time, the calculation changes from one of the yaw rate description of $\dot{\psi}_{\mathrm{T}}(t)$, $\dot{\psi}_{\mathrm{F}}(t)$, $\dot{\psi}_{\mathrm{R}}(t)$ to another one.

In Figure 13, all intersection times are shown. The relevant times for this example are written in bold.

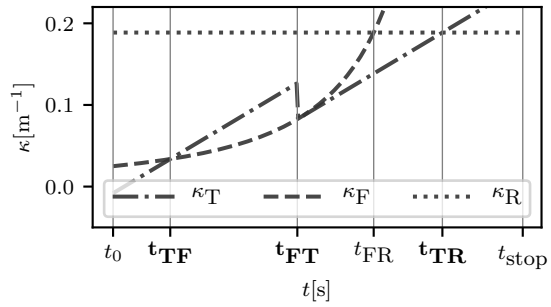


Figure 13. An example of relevant curvature intersections. All relevant intersections are marked in bold.

Depending on the input parameters, different types of trajectories evolve. In this paper, we distinguish between 9 different types. In Figure 14, we show all possible sub-elements of a trajectory and interpret the calculation of a trajectory as one of 9 different possible combinations of subsequent state changes.

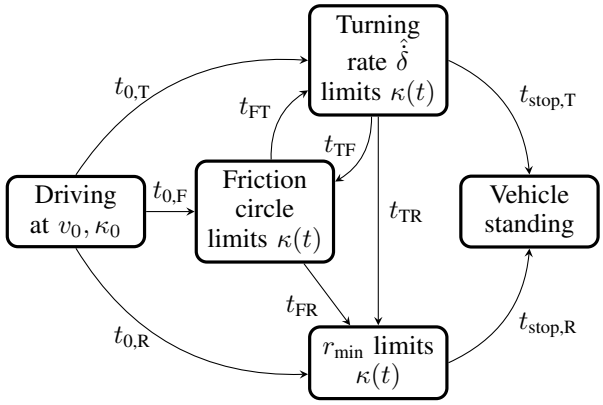


Figure 14. State diagram of all possible sub-elements of a trajectory. State changes occurs at intersections of curvatures and at t_0 and t_{stop}

Figure 14 shows which intersection times are relevant in each state. It also serves as a reference to find all possible combinations of states, and therefore all trajectory types that follow from the Extended Model.

Trajectory Type A: The state changes for Type A Trajectories are defined by the sequence $t_{0,\mathrm{T}} \to t_{\mathrm{TF}} \to t_{\mathrm{FR}} \to t_{\mathrm{stop,R}}$ in Figure 14.

Figure 15 shows the yaw rates of a Type A Trajectory. The effective yaw rate is always defined by the *minimum of all yaw rates* when turning left ($s = +$). This is marked by a thick line.

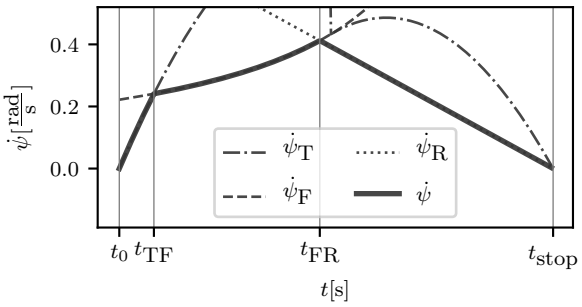


Figure 15. The yaw rates of a Type A trajectory. The effective yaw rate is marked by a thick line style.

The yaw angle function for all Type A trajectories is defined as Equation (73):

$$\psi(t) = \begin{cases} \overline{\psi_{\mathrm{T}}}(t) = \widetilde{\overline{\psi_{\mathrm{T}}}}(t) + C_{\psi,T}, & 0 \le t \le t_{\mathrm{TF}} \\ \psi_{\mathrm{F}}(t) = \widetilde{\psi_{\mathrm{F}}}(t) + C_{\psi,F}, & t_{\mathrm{TF}} < t \le t_{\mathrm{FR}} \\ \psi_{\mathrm{R}}(t) = \widetilde{\psi_{\mathrm{R}}}(t) + C_{\psi,R}, & t_{\mathrm{FR}} < t < t_{\mathrm{stop}} \end{cases} \tag{73}$$

where the constants of integration must be chosen specific to this trajectory type as shown in Equations (74) to (76):

$$C_{\psi,T} = \psi_0 \tag{74}$$

$$C_{\psi,F} = \overline{\psi_{\mathrm{T}}}(t_{\mathrm{TF}}) - \widetilde{\psi_{\mathrm{F}}}(t_{\mathrm{TF}}) \tag{75}$$

$$C_{\psi,R} = \psi_{\mathrm{F}}(t_{\mathrm{FR}}) - \widetilde{\psi_{\mathrm{R}}}(t_{\mathrm{FR}}) \tag{76}$$

The constants for $\Psi(t)$ are chosen in a way that all trajectory segments continue seamlessly.

The positions are calculated respectively in the same time intervals as in Equation (77):

$$p(t) = \begin{cases} p_{\mathrm{T}}(t) = \widetilde{p_{\mathrm{T}}}(t) + C_{\mathrm{p,T}}, & 0 \le t \le t_{\mathrm{TF}} \\ p_{\mathrm{F}}(t) = \widetilde{p_{\mathrm{F}}}(t) + C_{\mathrm{p,F}}, & t_{\mathrm{TF}} < t \le t_{\mathrm{FR}} \\ p_{\mathrm{R}}(t) = \widetilde{p_{\mathrm{R}}}(t) + C_{\mathrm{p,R}}, & t_{\mathrm{FR}} < t < t_{\mathrm{stop}} \end{cases} \tag{77}$$

where the constants of integration must be chosen in a way that all trajectory segments fit together seamlessly, as noted in Equations (78) to (80):

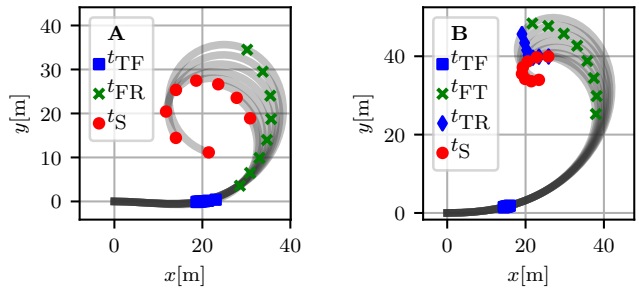
$$C_{p,T} = p_0 - \widetilde{p_{\mathrm{T}}}(0) \tag{78}$$

$$C_{p,F} = p_{\mathrm{T}}(t_{\mathrm{TF}}) - \widetilde{p_{\mathrm{F}}}(t_{\mathrm{TF}}) \tag{79}$$

$$C_{p,R} = p_{\mathrm{R}}(t_{\mathrm{FR}}) - \widetilde{p_{\mathrm{R}}}(t_{\mathrm{FR}}) \tag{80}$$

See Figure 16a for a set of different Type A Trajectories. The model parameters and inputs are equal in all examples of Type A, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type A Trajectories. Note that these trajectories are parametrized in a way that demonstrates the differences well.

Trajectory Type B: These trajectories are the most complex ones that can be described by the Extended Model, as it has the most different segment types. Examples are shown in Figure 16b. The resulting yaw rate segments are defined by the state change sequence $t_{0,\mathrm{T}} \to t_{\mathrm{TF}} \to t_{\mathrm{FT}} \to t_{\mathrm{TR}} \to t_{\mathrm{stop,R}}$ in Figure 14. Figure 17 shows the yaw rates of a Type B Trajectory, where the effective yaw rate is marked by a thick line.



(a) Eight Type A Trajectories. (b) Eight Type B Trajectories.

Figure 16. Type A and Type B trajectories. Both types are shown in rather theoretic situations for demonstration purposes.

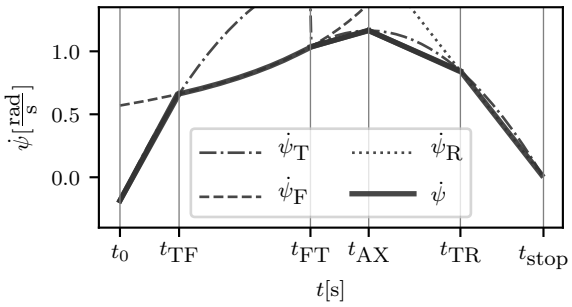


Figure 17. The yaw rates of a Type B trajectory. The five effective yaw rate segments are drawn bold, including the Apex Point of $\dot{\psi}_T$.

The yaw angle function for all Type B trajectories is defined as Equation (81):

$$\psi(t) = \begin{cases} \overline{\psi_{T,0}}(t) = \widetilde{\overline{\psi_T}}(t) + C_{\psi,T,0}, & 0 \le t \le t_{TF} \\ \psi_F(t) = \widetilde{\psi_F}(t) + C_{\psi,F}, & t_{TF} < t \le t_{FT} \\ \overline{\psi_{T,1}}(t) = \widetilde{\overline{\psi_T}}(t) + C_{\psi,T,1}, & t_{FT} < t \le t_{TR} \\ \psi_R(t) = \widetilde{\psi_R}(t) + C_{\psi,R}, & t_{TR} < t < t_{stop} \end{cases} \quad (81)$$

where the constants of integration must be chosen specific to this trajectory type as shown in Equations (82) to (85):

$$C_{\psi,T,0} = \psi_0 \quad (82)$$
$$C_{\psi,F} = \overline{\psi_T}(t_{TF}) - \widetilde{\psi_F}(t_{TF}) \quad (83)$$
$$C_{\psi,T,1} = \psi_F(t_{FT}) - \widetilde{\overline{\psi_T}}(t_{FT}) \quad (84)$$
$$C_{\psi,R} = \overline{\psi_T}(t_{TR}) - \widetilde{\psi_R}(t_{TR}) \quad (85)$$

The constants for $\Psi(t)$ are chosen in a way that all trajectory segments continue seamlessly.

The positions are calculated respectively in the same time intervals as in Equation (77):

$$p(t) = \begin{cases} p_{T,0}(t) = \widetilde{p_T}(t) + C_{p,T,0}, & 0 \le t \le t_{TF} \\ p_F(t) = \widetilde{p_F}(t) + C_{p,F}, & t_{TF} < t \le t_{FT} \\ p_{T,1}(t) = \widetilde{p_T}(t) + C_{p,T,1}, & t_{FT} < t \le t_{TR} \\ p_R(t) = \widetilde{p_R}(t) + C_{p,R}, & t_{TR} < t < t_{stop} \end{cases} \quad (86)$$

where the constants of integration must be chosen in a way that all trajectory segments fit together seamlessly, as noted in Equations (87) to (90):

$$C_{p,T,0} = p_0 - \widetilde{p_T}(0) \quad (87)$$
$$C_{p,F} = p_{T,0}(t_{TF}) - \widetilde{p_F}(t_{TF}) \quad (88)$$
$$C_{p,T,1} = p_F(t_{FT}) - \widetilde{p_{T,1}}(t_{FT}) \quad (89)$$
$$C_{p,R} = p_{T,1}(t_{TR}) - \widetilde{p_R}(t_{TR}) \quad (90)$$

See Figure 16b for a set of different Type B Trajectories. The model parameters and inputs are equal in all examples of Type B, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type B Trajectories.

Trajectory Type C: The state changes for Type C Trajectories are defined by the sequence $t_{0,T} \rightarrow t_{TR} \rightarrow t_{stop,R}$ in Figure 14.
Figure 18 shows the yaw rates of a Type C Trajectory. The effective yaw rate is marked by a thick line.

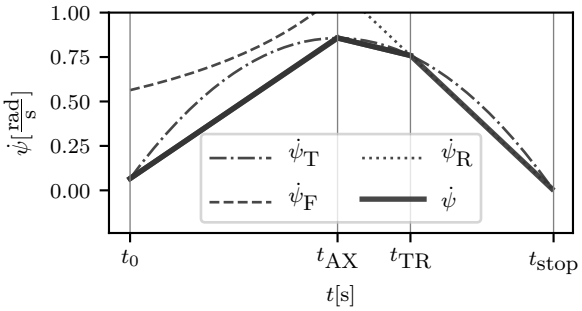


Figure 18. The yaw rates of a Type C trajectory. The effective yaw rate is marked by a thick line style.

The yaw angle function for all Type C trajectories is defined as Equation (91):

$$\psi(t) = \begin{cases} \overline{\psi_T}(t) = \widetilde{\overline{\psi_T}}(t) + C_{\psi,T}, & 0 \le t \le t_{TR} \\ \psi_R(t) = \widetilde{\psi_R}(t) + C_{\psi,R}, & t_{TR} < t < t_{stop} \end{cases} \quad (91)$$

where the constants of integration must be chosen specific to this trajectory type as shown in Equations (92) and (93):

$$C_{\psi,T} = \psi_0 \quad (92)$$
$$C_{\psi,R} = \psi_T(t_{TR}) - \widetilde{\psi_R}(t_{TR}) \quad (93)$$

The constants for $\Psi(t)$ are chosen in a way that all trajectory segments continue seamlessly.

The positions are calculated respectively in the same time intervals as in Equation (94):

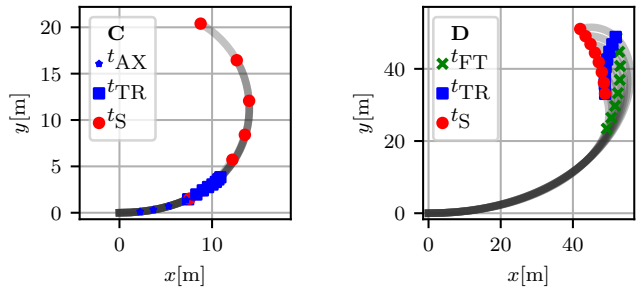
$$p(t) = \begin{cases} p_T(t) = \widetilde{p_T}(t) + C_{p,T}, & 0 \le t \le t_{TR} \\ p_R(t) = \widetilde{p_R}(t) + C_{p,R}, & t_{TR} < t < t_{stop} \end{cases} \quad (94)$$

where the constants of integration must be chosen in a way that all trajectory segments fit together seamlessly, as noted in Equations (95) and (96):

$$C_{p,T} = p_0 - \widetilde{p_T}(0) \quad (95)$$
$$C_{p,R} = p_R(t_{TR}) - \widetilde{p_R}(t_{TR}) \quad (96)$$

See Figure 19a for a set of different Type C Trajectories. The model parameters and inputs are equal in all examples of Type C, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type C Trajectories. It can be seen that all trajectories are on top of each other, which

makes sense, because the braking factor only has influence on the actual direction when the Friction Circle is involved in the calculation. In Type C trajectories, this is not the case.



(a) Eight Type C Trajectories. (b) Eight Type D Trajectories.

Figure 19. Different trajectories of both Type C and Type D.

Trajectory Type D: The state changes for Type D Trajectories are defined by the sequence $t_{0,F} \rightarrow t_{FT} \rightarrow t_{TR} \rightarrow t_{stop,R}$ in Figure 14. Hence, Type D Trajectories are like Type B but are limited by the Friction Circle from the beginning on.

Figure 20 shows the yaw rates of a Type D Trajectory. The effective yaw rate is always defined by the *minimum of all yaw rates* when turning left ($s = +$). This is marked by a thick line in Figure 20.

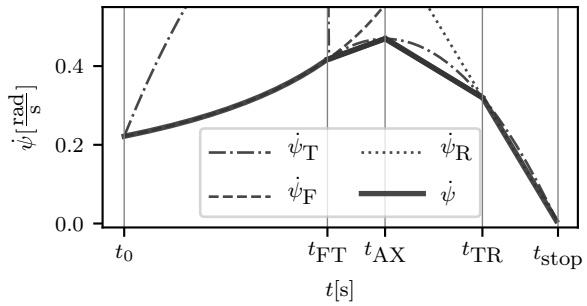


Figure 20. The yaw rates of a Type D trajectory. The effective yaw rate is marked by a thick line style.

The yaw angle function for all Type D trajectories is defined as Equation (97):

$$\psi(t) = \begin{cases} \psi_F(t) = \widetilde{\psi_F}(t) + C_{\psi,F}, & 0 < t \leq t_{FT} \\ \widetilde{\psi_T}(t) = \widetilde{\psi_T}(t) + C_{\psi,T}, & t_{FT} < t \leq t_{TR} \\ \psi_R(t) = \widetilde{\psi_R}(t) + C_{\psi,R}, & t_{TR} < t < t_{stop} \end{cases} \quad (97)$$

where the constants of integration must be chosen specific to this trajectory type as shown in Equations (98) to (100):

$$C_{\psi,F} = \psi_0 \quad (98)$$
$$C_{\psi,T} = \psi_F(t_{TF}) - \widetilde{\psi_F}(t_{TF}) \quad (99)$$
$$C_{\psi,R} = \psi_T(t_{TR}) - \widetilde{\psi_R}(t_{TR}) \quad (100)$$

The constants for $\Psi(t)$ are chosen in a way that all trajectory segments continue seamlessly.

The positions are calculated respectively in the same time intervals as in Equation (101):

$$p(t) = \begin{cases} p_F(t) = \widetilde{p_F}(t) + C_{p,F}, & 0 \leq t \leq t_{FT} \\ p_T(t) = \widetilde{p_T}(t) + C_{p,T}, & t_{FT} < t \leq t_{TR} \\ p_R(t) = \widetilde{p_R}(t) + C_{p,R}, & t_{TR} < t < t_{stop} \end{cases} \quad (101)$$

where the constants of integration must be chosen in a way that all trajectory segments fit together seamlessly, as noted in Equations (102) to (104):

$$C_{p,F} = p_0 - \widetilde{p_F}(0) \quad (102)$$
$$C_{p,T} = p_T(t_{FT}) - \widetilde{p_T}(t_{FT}) \quad (103)$$
$$C_{p,R} = p_R(t_{TR}) - \widetilde{p_R}(t_{TR}) \quad (104)$$

See Figure 19b for a set of different Type D Trajectories. The model parameters and inputs are equal in all examples of Type D, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type D Trajectories.

Trajectory Type E: The state changes for Type E Trajectories are defined by the sequence $t_{0,F} \rightarrow t_{FR} \rightarrow t_{stop,R}$ in Figure 14.

Figure 21 shows the yaw rates of a Type E Trajectory. The effective yaw rate is marked by a thick line.

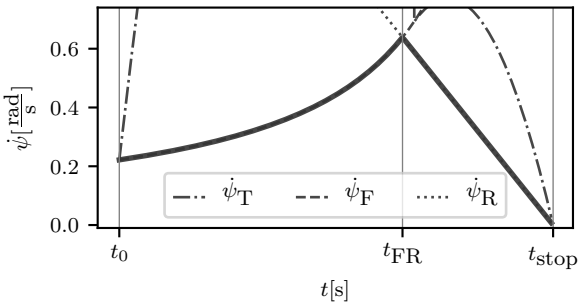


Figure 21. The yaw rates of a Type E trajectory. The effective yaw rate is marked by a thick line style.

The yaw angle function for all Type E trajectories is defined as Equation (105):

$$\psi(t) = \begin{cases} \psi_F(t) = \widetilde{\psi_F}(t) + C_{\psi,F}, & 0 \leq t \leq t_{FR} \\ \psi_R(t) = \widetilde{\psi_R}(t) + C_{\psi,R}, & t_{FR} < t < t_{stop} \end{cases} \quad (105)$$

where the constants of integration must be chosen specific to this trajectory type as shown in Equations (106) and (107):

$$C_{\psi,F} = \psi_0 \quad (106)$$
$$C_{\psi,R} = \psi_F(t_{FR}) - \widetilde{\psi_R}(t_{FR}) \quad (107)$$

The constants for $\Psi(t)$ are chosen in a way that all trajectory segments continue seamlessly.

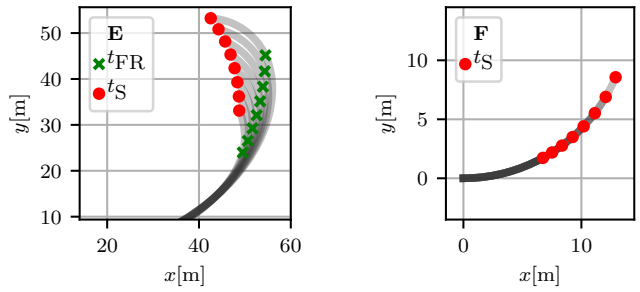
The positions are calculated respectively in the same time intervals as in Equation (108):

$$p(t) = \begin{cases} p_F(t) = \widetilde{p_F}(t) + C_{p,F}, & 0 \leq t \leq t_{FR} \\ p_R(t) = \widetilde{p_R}(t) + C_{p,R}, & t_{FR} < t < t_{stop} \end{cases} \quad (108)$$

where the constants of integration must be chosen in a way that all trajectory segments fit together seamlessly, as noted in Equations (109) and (110):

$$C_{p,F} = p_0 - \widetilde{p_F}(0) \quad (109)$$
$$C_{p,R} = p_R(t_{FR}) - \widetilde{p_R}(t_{FR}) \quad (110)$$

See Figure 22a for a set of different Type E Trajectories. The model parameters and inputs are equal in all examples of Type E, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type E Trajectories.



(a) Eight Type E Trajectories. (b) Eight Type F Trajectories.

Figure 22. Different trajectories of both Type E and Type F.

Trajectory Type F: The state changes for Type F Trajectories are defined by the simple sequence $t_{0,R} \rightarrow t_{stop,R}$ in Figure 14. Figure 23 shows the yaw rate of a Type F Trajectory. The effective yaw rate is marked by a thick line to underline the simplicity of Type F trajectories.

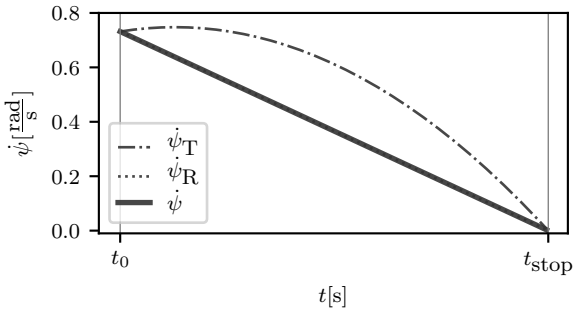


Figure 23. The yaw rates of a Type F trajectory. The effective yaw rate is drawn bold. Note that in this Type $\dot{\psi}$ and $\dot{\psi}_R$ are identical.

The yaw angle function for all Type F trajectories is defined as Equation (111):

$$\psi(t) = \psi_R(t) = \widetilde{\psi_R}(t) + C_{\psi,R}, \quad t_0 \leq t < t_{stop} \tag{111}$$

where the constant of integration must be chosen specific to this trajectory type as shown in Equation (112):

$$C_{\psi,R} = \psi_0 - \widetilde{\psi_R}(t_0) \tag{112}$$

The constant for $\Psi(t)$ is chosen in a way that the trajectory starts correctly with ψ_0.

The positions are calculated respectively in the same time intervals as in Equation (113):

$$p(t) = p_R(t) = \widetilde{p_R}(t) + C_{p,R}, \quad t_0 \leq t < t_{stop} \tag{113}$$

where the constant of integration must be chosen in a way that the trajectory starts correctly at p_0, as noted in Equation (114):

$$C_{p,R} = p_0 - \widetilde{p_R}(t_0) \tag{114}$$

See Figure 22b for a set of different Type F Trajectories. The model parameters and inputs are equal in all examples of Type F, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type F Trajectories. It can be seen that all trajectories are on top of each other, which makes sense, because the braking factor only has influence on the actual direction when the Friction Circle is involved in the calculation. In Type F trajectories this is not the case.

Trajectory Type G: Type G Trajectories are similar to Type B Trajectories, but in contrast they end before reaching r_{turn}. The segments are defined by the state change sequence $t_{0,T} \rightarrow t_{TF} \rightarrow t_{FT} \rightarrow t_{stop,T}$ in Figure 14. Figure 24 shows the yaw rates of a Type G Trajectory. The effective yaw rate is marked by a thick line in Figure 24.

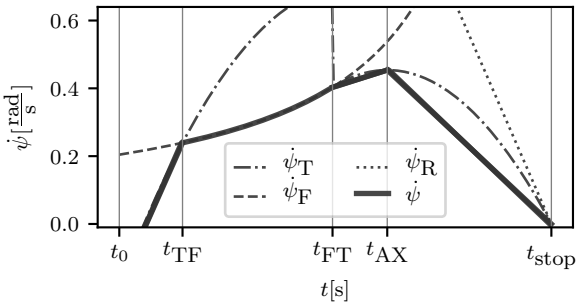


Figure 24. The yaw rates of a Type G trajectory. The effective yaw rate is marked by a thick line style.

The yaw angle function for all Type G trajectories is defined as Equation (115):

$$\psi(t) = \begin{cases} \overline{\psi_{T,0}}(t) = \widetilde{\psi_T}(t) + C_{\psi,T,0}, & 0 \leq t \leq t_{TF} \\ \psi_F(t) = \widetilde{\psi_F}(t) + C_{\psi,F}, & t_{TF} < t \leq t_{FT} \\ \overline{\psi_{T,1}}(t) = \widetilde{\psi_T}(t) + C_{\psi,T,1}, & t_{FT} < t \leq t_{stop} \end{cases} \tag{115}$$

where the constants of integration must be chosen specific to this trajectory type as shown in Equations (116) to (118):

$$C_{\psi,T,0} = \psi_0 \tag{116}$$

$$C_{\psi,F} = \overline{\psi_T}(t_{TF}) - \widetilde{\psi_F}(t_{TF}) \tag{117}$$

$$C_{\psi,T,1} = \psi_F(t_{FT}) - \widetilde{\overline{\psi_T}}(t_{FT}) \tag{118}$$

The constants for $\Psi(t)$ are chosen in a way that all trajectory segments continue seamlessly.

The positions are calculated respectively in the same time intervals as in Equation (119):

$$p(t) = \begin{cases} p_{T,0}(t) = \widetilde{p_T}(t) + C_{p,T,0}, & 0 \leq t \leq t_{TF} \\ p_F(t) = \widetilde{p_F}(t) + C_{p,F}, & t_{TF} < t \leq t_{FT} \\ p_{T,1}(t) = \widetilde{p_T}(t) + C_{p,T,1}, & t_{FT} < t \leq t_{stop} \end{cases} \tag{119}$$

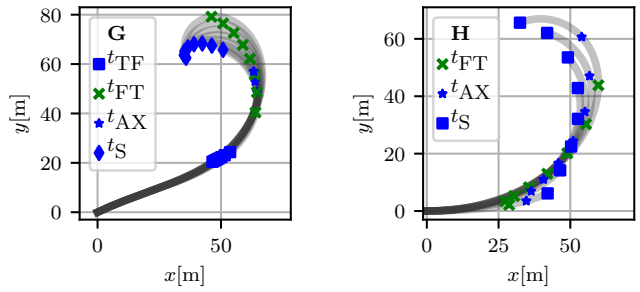
where the constants of integration must be chosen in a way that all trajectory segments fit together seamlessly, as noted in Equations (120) to (122):

$$C_{p,T,0} = p_0 - \widetilde{p_T}(0) \tag{120}$$

$$C_{p,F} = p_{T,0}(t_{TF}) - \widetilde{p_F}(t_{TF}) \tag{121}$$

$$C_{p,T,1} = p_F(t_{FT}) - \widetilde{p_{T,1}}(t_{FT}) \tag{122}$$

See Figure 25a for a set of different Type G Trajectories. The model parameters and inputs are equal in all examples of Type G, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type G Trajectories.



(a) Eight Type G Trajectories. (b) Eight Type H Trajectories.

Figure 25. Different trajectories of both Type G and Type H.

Trajectory Type H: The state changes for Type H Trajectories are defined by the sequence $t_{0,F} \rightarrow t_{FT} \rightarrow t_{stop,R}$ in Figure 14.

Figure 26 shows the yaw rates of a Type H Trajectory. The effective yaw rate is marked by a thick line.

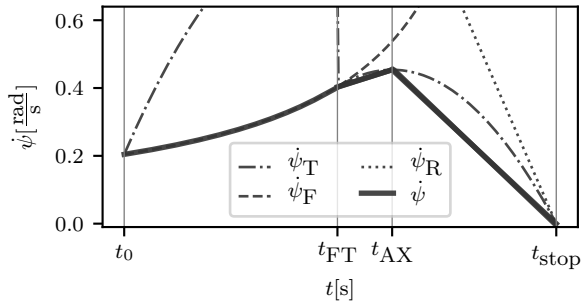


Figure 26. The yaw rates of a Type H trajectory. The effective yaw rate is drawn bold. Notice the similarity to Type G Trajectories as shown in Figure 24

The yaw angle function for all Type H trajectories is defined as Equation (123):

$$\psi(t) = \begin{cases} \psi_F(t) = \widetilde{\psi_F}(t) + C_{\psi,F}, & 0 \le t \le t_{FT} \\ \psi_T(t) = \widetilde{\psi_T}(t) + C_{\psi,T}, & t_{FT} < t < t_{stop} \end{cases} \quad (123)$$

where the constants of integration must be chosen specific to this trajectory type as shown in Equations (124) and (125):

$$C_{\psi,F} = \psi_0 \quad (124)$$

$$C_{\psi,T} = \psi_F(t_{FT}) - \widetilde{\psi_R}(t_{FT}) \quad (125)$$

The constants for $\Psi(t)$ are chosen in a way that all trajectory segments continue seamlessly.

The positions are calculated respectively in the same time intervals as in Equation (126):

$$p(t) = \begin{cases} p_F(t) = \widetilde{p_F}(t) + C_{p,F}, & 0 \le t \le t_{FT} \\ p_T(t) = \widetilde{p_T}(t) + C_{p,T}, & t_{FT} < t < t_{stop} \end{cases} \quad (126)$$

where the constants of integration must be chosen in a way that all trajectory segments fit together seamlessly, as noted in Equations (127) and (128):

$$C_{p,F} = p_0 - \widetilde{p_F}(0) \quad (127)$$

$$C_{p,T} = p_T(t_{FT}) - \widetilde{p_R}(t_{FT}) \quad (128)$$

See Figure 25b for a set of different Type H Trajectories. The model parameters and inputs are equal in all examples of Type H, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type H Trajectories.

Trajectory Type I: The state changes for Type I Trajectories are defined by the simple sequence $t_{0,T} \rightarrow t_{stop,R}$ in Figure 14.

Figure 27 shows the yaw rate of a Type I Trajectory. The effective yaw rate is marked by a thick line to underline the simplicity of Type I trajectories.

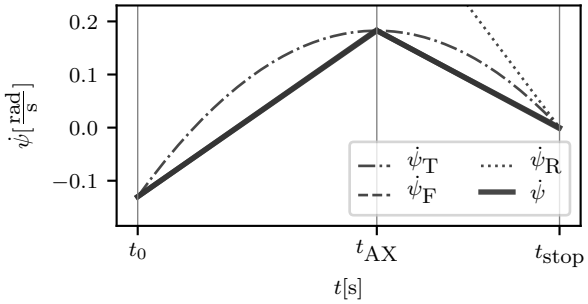


Figure 27. The yaw rates of a Type I trajectory. The effective yaw rate is marked by a thick line style.

The yaw angle function for all Type I trajectories is defined as Equation (129):

$$\psi(t) = \overline{\psi_T}(t) = \widetilde{\overline{\psi_T}}(t) + C_{\psi,T}, \quad t_0 \le t < t_{stop} \quad (129)$$

where the constant of integration must be chosen specific to this trajectory type as shown in Equation (130):

$$C_{\psi,T} = \psi_0 - \widetilde{\psi_T}(t_0) \quad (130)$$

The constant for $\Psi(t)$ is chosen in a way that the trajectory starts correctly with ψ_0.

The positions are calculated respectively in the same time intervals as in Equation (131):

$$p(t) = p_T(t) = \widetilde{p_T}(t) + C_{p,T}, \quad t_0 \le t < t_{stop} \quad (131)$$

where the constant of integration must be chosen in a way that the trajectory starts correctly at p_0, as noted in Equation (132):

$$C_{p,T} = p_0 - \widetilde{p_T}(t_0) \quad (132)$$

See Figure 28 for a set of different Type I Trajectories. The model parameters and inputs are equal in all examples of Type I, except for the Braking Factor b, which is chosen to be in an admissible range to result in Type I Trajectories. It can be seen that all trajectories are on top of each other, which makes sense, because the braking factor only has influence on the actual direction when the Friction Circle is involved in the calculation. In Type I trajectories this is not the case.

V. Conclusion

In this paper, we present two models for hard braking and collision avoiding vehicle trajectories. In the Basic Model, we take into account the maximally applicable acceleration/deceleration between tires and road surface, the minimal turning radius, the vehicle velocity, as well as starting position and heading. We explain our approach in detail and compare our model equations with an iterative CTRA-Model

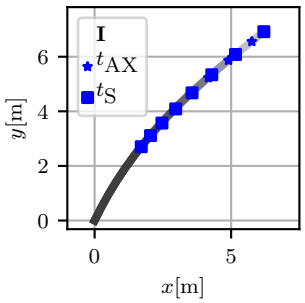


Figure 28. Eight trajectories of Type I. Note that since the Friction Circle has no influence in these trajectories, they are all located along the same path.

simulation, which finds very similar solutions. However, in tests we could show that our Basic Model computes stopping positions and trajectories up to 20 times faster than CTRA. The Extended Model furthermore take into account the initial yaw rate of a vehicle, as well as its ability to change the steering angle. In order to translate the steering angle into a yaw rate, we introduce the length of a vehicle and its turning direction. The increased complexity leads to 9 different trajectory types, which we explore in depth. By solving the compound differential equations of both models for position in x, y-plane, we describe the complete vehicle motion till full stop, while also turning and still respecting the Friction Circle. With the derived equations, we can directly compute possible positions that a vehicle will reach in a braking and collision avoiding scenario. This might be used to generate braking and collision avoiding trajectories, by sampling feasible motion primitives, which can be computed in very short time.

We contribute a Basic Model, which can aid in solving reachability problems for hard braking vehicles in an accurate and yet overapproximative way. Furthermore we contribute a second, Extended Model, which takes into account initial vehicle dynamics and respects further physical constraints. The second model is therefore much better suited for calculating motion primitives of actual emergency trajectories.

As next steps, the proposed models for vehicle motion can be compared to the trajectories of real vehicles under the same assumptions given. Another next step might be the usage of our model for fast generation of braking trajectories by sampling motion primitives and a comparison to other state of the art methods. We suggest to search a tree structure of connected motion primitives sampled from the Extended Model for feasible and yet complex evasive trajectories. As we can directly compute motion primitives for the highly non linear motions in braking and collision avoidance, the proposed model can significantly reduce valuable calculation time. Another application is the application of both models in a formal reachability analysis for risk assessment in hard braking traffic scenarios and compare the solution to other contributions in the field of reachability analysis.

References

[1] F. Terhar and C. Icking, "A New Model for Hard Braking Vehicles and Collision Avoiding Trajectories," in *Proceedings of the 8th International Conference on Advances in Vehicular Systems, Technologies and Applications*, June 2019, pp. 28–33.

[2] M. Werling, J. Ziegler, S. Kammel, and S. Thrun, "Optimal trajectory generation for dynamic street scenarios in a frenet frame," in *Proceedings - IEEE International Conference on Robotics and Automation*, 06 2010, pp. 987 – 993.

[3] J. Ziegler, M. Werling, and J. Schröder, "Navigating car-like robots in unstructured environments using an obstacle sensitive cost function," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 07 2008, pp. 787 – 791.

[4] C. Pek and M. Althoff, "Computationally efficient fail-safe trajectory planning for self-driving vehicles using convex optimization," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2018, pp. 1447–1454.

[5] S. Magdici and M. Althoff, "Fail-safe motion planning of autonomous vehicles," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2016, pp. 452–458.

[6] C. Pek, M. Koschi, and M. Althoff, "An online verification framework for motion planning of self-driving vehicles with safety guarantees," in *AAET - Automatisiertes und vernetztes Fahren*, 01 2019, pp. 260–274.

[7] S. Heinrich, "Planning Universal On-Road Driving Strategies for Automated Vehicles." "Springer Fachmedien Wiesbaden", 2018, ch. 4, pp. 33–47.

[8] I. M. Mitchell, "Comparing forward and backward reachability as tools for safety analysis," in *Hybrid Systems: Computation and Control*, A. Bemporad, A. Bicchi, and G. Buttazzo, Eds., 2007, pp. 428–443.

[9] E. Asarin, T. Dang, and A. Girard, "Reachability analysis of nonlinear systems using conservative approximation," in *Hybrid Systems: Computation and Control*, O. Maler and A. Pnueli, Eds., 2003, pp. 20–35.

[10] A. Girard, "Reachability of uncertain linear systems using zonotopes," in *Hybrid Systems: Computation and Control*, M. Morari and L. Thiele, Eds., 2005, pp. 291–305.

[11] M. Koschi and M. Althoff, "SPOT: A tool for set-based prediction of traffic participants," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 1686–1693.

[12] M. Althoff, "Reachability analysis and its application to the safety assessment of autonomous cars," Dissertation, Technische Universität München, München, 2010.

[13] B. Kim *et al.*, "Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network," *CoRR*, vol. abs/1704.07049, 2017.

[14] F. Giovannini, G. Savino, and M. Pierini, "Influence of the minimum swerving distance on the development of powered two wheeler active braking," in *22nd ESV Conference*, June 2011, paper 11-0258.

[15] C. Ackermann, J. Bechtloff, and R. Isermann, "Collision avoidance with combined braking and steering," in *6th International Munich Chassis Symposium 2015*, P. Pfeffer, Ed., 2015, pp. 199–213.

[16] C. Choi and Y. Kang, "Simultaneous braking and steering control method based on nonlinear model predictive control for emergency driving support," *International Journal of Control, Automation and Systems*, vol. 15, no. 1, pp. 345–353, Feb 2017.

[17] J. Stewart, "Calculus: Early transcendentals," L. Covello, Ed. Thomson Brooks/Cole, 2010, ch. 5.

[18] H. B. Pacejka, "Chapter 1 - Tire characteristics and vehicle handling and stability," in *Tire and Vehicle Dynamics (Third edition)*, H. B. Pacejka, Ed., 2012.

[19] R. Schubert, E. Richter, and G. Wanielik, "Comparison and evaluation of advanced motion models for vehicle tracking," in *11th International Conference on Information Fusion*, June 2008, pp. 1–6.

[20] S. Söntges and M. Althoff, "Computing the drivable area of autonomous road vehicles in dynamic road scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 6, pp. 1855–1866, June 2018.

[21] "*NIST Digital Library of Mathematical Functions*," http://dlmf.nist.gov/, Release 1.0.25 of 2019-12-15, f. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[22] I. Stewart, "Taming the Infinite: The Story of Mathematics from the First Numbers to Chaos Theory." Quercus, 2008, ch. 4.

Age-Related Differences of Cognitive Functions when Encountering Driving Hazards on Expressways

Kazuhito Sato, Yuki Oomomo, Hirokazu Madokoro
Department of Machine Intelligence and Systems Engineering,
Faculty of Systems Science and Technology, Akita Prefectural University
Yurihonjo, Japan
Email: ksato, m18a016, madokoro @akita-pu.ac.jp

Momoyo Ito
Institute of Technology and Science,
Tokushima University
Tokushima, Japan
Email: momoito@is.tokushima-u.ac.jp

Sakura Kadowaki
SmartDesign Co., Ltd.
Akita, Japan
Email: sakura@smart-d.jp

Abstract - **For this study, we defined four situations that are likely to trigger accidents on highways as driving hazards: vehicle breakdowns, sudden appearance of small animals, falling objects, and lane decreases. Carrying out running experiments while controlling the time of day (daytime or nighttime) and traffic flow (presence or absence of overtaking vehicle), we conducted a comparison for each driving topic, particularly addressing driving characteristics of young drivers and elderly drivers related to "cognition", "judgment", and "operations". Using evaluation experiment results, we strove to analyze the driving characteristics of elderly drivers for driving hazards encountered on highways based on comparison to those of young drivers. Results revealed findings unique to elderly drivers for response after "cognition" and reaction times necessary for "judgment and operations".**

Keywords – driving behavior; driving topic; elderly drivers; highway; human engineering

I. INTRODUCTION

Although the annual numbers of deaths caused by traffic accidents in Japan show a downward trend, traffic accidents caused by erroneous operations by elderly drivers and reverse running on highways have become particularly problematic in recent years. Traffic accidents caused by elderly drivers are frequent and entail a high risk of severe injury. Therefore, countermeasures must be taken to cope with Japan's impending super aging society.

As factors contributing to traffic accidents by elderly drivers, there might be common points among elderly people such as slowed driving operations related to cognition, judgment, and operations, mistaken operations of handles and brakes because of the distribution of attention to multiple tasks and concentration, and momentary carelessness and distraction. Recently, various traffic accident preventive safety systems have been studied, as represented by automatic driving efforts. Practical application of these systems has been accelerating. However, most such systems merely collect and analyze external information of the vehicle. Such systems cannot deal adaptively with the driving characteristics of individual drivers. Furthermore, preventive safety technologies based on some standard general driver make it difficult to address the slowing of driving behaviors that are peculiar to elderly people, related to recognition, judgment, and operations. Approaches must be made for preventive safety systems dedicated to the driving characteristics of elderly drivers.

Our earlier study [1] examined four driving situations as driving hazards that are likely to trigger accidents: vehicle breakdowns on the road shoulder, sudden appearance of small animals, falling objects, and lane decreases. Those situations are especially dangerous for highways: they have high mortality rates and easily lead to severe accidents. We conducted running experiments while controlling the time of day (daytime or nighttime) and traffic flow (with and without overtaking vehicles). Specifically, we devoted attention to a series of driving behaviors related to cognition, judgment, and operations for the respective driving topics. We compared and analyzed the driving characteristics of young drivers and elderly drivers. Results clarified that elderly drivers had longer reaction times than young drivers. Particularly, the reaction time associated with "cognition" was remarkable. However, we have not validated the findings. The running scenarios were limited for analyses in terms of the time of day (daytime) and the traffic flow (no overtaking vehicle), with few elderly driver participants.

This study clarifies the driving characteristics of elderly drivers for these driving hazards by increasing the number of elderly driver participants and performing comparison with data of young drivers, examining influences of visibility because of changes in the time of day (daytime / nighttime) and changes in the surrounding cognitive situation because of the presence or absence of overtaking vehicles.

This paper is presented as follows. We review related work to clarify the position of this study among existing studies of

the literature in Section II. Section III presents a definition of the experiment protocols, driving topics, and running scenarios. In Section IV, we examine the correspondence relation between measurement points of driver reaction and eye-gaze information, particularly addressing the reaction time necessary for each action, i.e., cognition, judgment, and operations for driving topics. Additionally, we compare the average reaction times of all drivers, elderly drivers, and young drivers at each measurement point for "falling objects", which is one of the driving topics. Finally, we present conclusions and intentions for future work in Section V.

II. Related Works

Causes of fatal accidents involving are mostly unsuitable driving operations and are attributed to physical characteristics, psychological characteristics, driving characteristics, and social characteristics of the elderly drivers themselves. Physical characteristics indicate declining physical function such as vision and exertion ability. Assistive technologies and systems can reduce the operation load imposed by steering. Some systems automatically reduce the speed and mitigate damage when a possibility of a collision arises. Many systems have been put into practical use [2]. Psychological characteristics of elderly drivers render them as less able at parallel processing of multiple streams of information. They tend to become self-oriented. Driving characteristics show a mismatch of consciousness and behaviors caused by feeling overly "accustomed" to driving or relying on skills of "driving that would be" based on past experiences. Additionally, social characteristics imply differences in characteristics by generations, such as a decline in communication skills and the influence of automation [3]. As solutions to these characteristics which might impede elderly drivers, the following have been reported. It is possible to reform consciousness and motivate people for safe driving by having each elderly person participate in efforts for traffic safety by their own will rather than passively [3]. Elderly drivers are able to improve their attention abilities by pointing and designating signs to recognize [4].

Recent studies examining elderly drivers have been specifically undertaken by the Nagoya COI project, which was adopted by the Ministry of Education, Culture, Sports, Science and Technology as an innovative creation program [5]. It has pursued the construction of a human aging driving characteristics database (Dahlia) promoted as part of this project, with cognitive function tests, driving aptitude tests, visual function tests, driving characteristics surveys, etc. They have been conducted for 300 elderly people each year [6]. The project collects and analyzes actual road driving data from a drive recorder and driving test data from a driving simulator. The results clarify correlation between cognitive and visual functions of elderly people and collision rates [7]. Additionally, a driver agent system was developed to encourage and improve elderly drivers' good driving behavior [8] [9]. Experiments were conducted to evaluate the acceptability of the system. Results demonstrated the effectiveness of repeatedly experiencing video feedback [10] [11].

Furthermore, Takahara et al. [12] analyzed the characteristics at a temporary stop location to study the cognitive function of elderly drivers. Results demonstrated the effectiveness of their system through development of a voice guidance type temporary stop support system. Iida et al. [13] assessed a hypothesis of reverse running processes as an example of elderly drivers' accidents on an expressway. They confirmed that the psychological state of the elderly driver and the road composition make reverse running occur easily.

Abe et al. [14] constructed a driving simulator experiment to investigate driving condition and traffic environment effects on visual behavior. Specifically, they examined how a visual field is influenced as a function of cognitive distraction. They simulated cognitive distraction by an experimental secondary task related to mental calculations. Their results indicated that a driver's reaction time to a target mark was increased as a result of cognitive distraction. Honma et al. [15] examined how drivers' visual attention is influenced when a driver is drowsy. They used the knowledge to assess driver impairment related to object recognition. Driving simulator experiment results showed that drowsy driving impaired drivers' visual attention, particularly on conditions with no vehicle ahead, compared to normal driving.

For research and development of Autonomous Driving Systems (i.e., Strategic Innovation Creation Program: SIP), which have recently attracted much attention, legal systems and infrastructure developments have been promoted to realize automatic driving on expressways and autonomous automatic driving in restricted areas [16]. Wada et al. [17] compared and verified perceptions of elderly drivers and general drivers on the expressway during autonomous driving based on SAE level 2 [18]. As in their earlier studies, results of analyzing characteristics of elderly drivers using a semi-autonomous vehicle (i.e., pro-pilot) of SAE Level 2 confirmed that peripheral cognitive levels decrease [19] [20].

Nevertheless, these studies have not particularly analyzed situations (driving topics or hazards) that readily induce accidents on highways, such as vehicle breakdowns on the road shoulder, sudden movements of small animals onto the roadway, falling objects, and reduced lanes during maintenance work. This study used driving experiments with controlled time of day and traffic flow to analyze driving behaviors related to cognition, judgment, and operations of these driving topics. Results clarify the elderly drivers' driving characteristics.

III. Methods

This section defines the experiment protocols and four driving hazards that are likely to cause accidents. Next, we present running scenarios with controlled daytime or nighttime, and presence or absence of an overtaking vehicle.

A. Experiment Systems

Although many people perform driving behaviors every day, many difficulties arise in clarifying individual driving characteristics from actual behaviors on the road in real environments. Driving behaviors vary depending on the road environment and traffic conditions prevailing at the time. Actual road conditions cannot be reproduced constantly. If any behavior varies, distinguishing clearly whether the variation is caused by a difference in traffic conditions or by variation among individuals is not possible.

This study used a driving simulator (DS) to assess driving behaviors for freely set road environments and traffic conditions affecting driver behaviors. Figure 1 portrays the experiment system configured to measure driver behaviors. The DS used for experiments has platforms corresponding to compact and six-axis motion, which is equipped with ordinary cars. The DS has three-color liquid-crystal displays mounted at the cabin front. It also has a function reproducing pseudo-driving environments that are freely configurable to horizontal viewing angles. Figure 1 shows cameras installed to the left and right and center of the three-color liquid-crystal monitors mounted at the cabin front. Without restraining drivers, they gather body information such as head poses, face orientations, and eye-gaze movements. Additionally, we set an infrared pod on top of the instruments at the cabin front. The camera heads and infrared pod are input-based sensors of a head-gaze tracking device (FaceLAB; Ekstreme Machine Corp.). Through preliminary test runs with multiple subjects, we confirmed that the stereo camera head and the infrared pod installation do not interfere with visibility during driving. To capture driver facial expressions, a USB camera (Xtion pro Live; ASUS Corp.) was installed at the top of the liquid crystal monitor at the center of the cabin front.

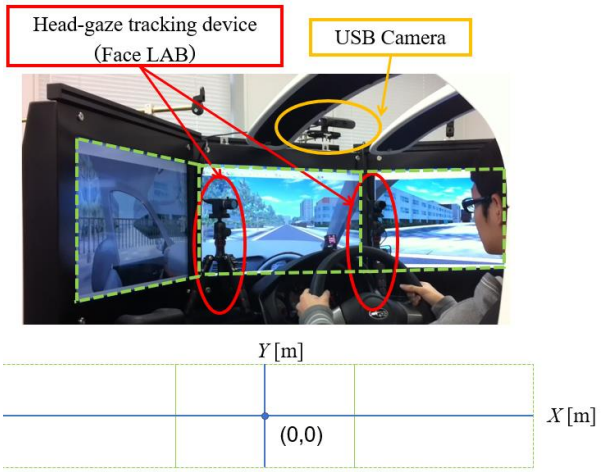


Figure 1. Experiment system for measuring driver behaviors.

B. Experiment Protocols

Figure 2 presents the experiment protocol outline. Initially, as individual characteristics of each subject, we conducted an examination of the following questionnaire methods using the driving style check sheet: attitude, orientation, and concept to work on driving [21]. The operation burdens were imposed using a driving load sensitivity check sheet [22]. For one running test, each target subject wore a heart rate monitor (RS800CX; Polar). We measured the instantaneous heart rate during a normal state at 1 min in advance. Next, to improve the measurement accuracy for face orientation and eye-gaze movements of each participant, we calibrated the cameras for use with a head-gaze tracking device (FaceLAB). We recorded a video of the driver's face while driving with the USB camera (Xtion Pro Live; ASUS Corp.) to analyze the driver facial expressions. After these preparations, each participant ran along the four running scenarios described in Section III.C by synchronizing the time bases of all measuring devices. Finally, a questionnaire specifically asking about the driving hazards given at random, subjective reviews, a four-stage check was administered when a driving hazard occurred. After obtaining approval of the Akita

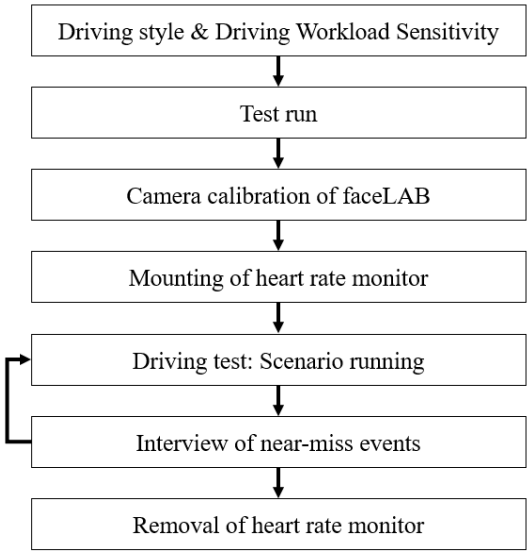
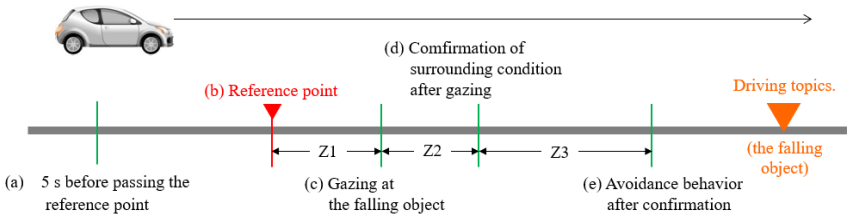


Figure 2. Outline of experiment protocols.

Prefectural University Research Ethics Board, the experiment contents were explained completely to participants, after which we obtained written consent from participants. We also obtained their agreement to publish their face images. Participants were 26 people: 10 young drivers (avg. 22 years old) and 16 elderly drivers (avg. 64 years old).

C. Driving hazards and Running scenarios

The driving course used for the experiment is a straight course of two lanes with no lane changes on one side simulating the Tohoku Expressway. The course starts running from the driving scene, which converges from the acceleration lane to the expressway. Driving hazards in one



	(a)	(b)	(c)	(d)	(e)
Velocity	101 km/h	100 km/h	98 km/h	98 km/h	96 km/h
Time(Video)	1:47.67	1:52.65	1:55.93	1:56.26	1:58.63
		Z1 3.28 s	Z2 0.33 s	Z3 2.37 s	

Figure 3. Measurement points of driver reaction to driving topics.

run were set randomly. Measurement points of each driver's reaction for each driving hazard are shown in Figure 3. Figure 4 presents the state of each driving hazard. The falling object in Figure 4(a) is installed at the center of the left lane. It interferes with the running course. The sudden appearance of small animals shown in Figure 4(b) was set so that when the vehicle passed the fixed point, it emerged from the front left side of the vehicle and traversed to the right to close the driving path of the vehicle. The broken down vehicle in Figure 4(c) represented a large truck parked on the shoulder because of a failure. Because of the lane decrease in Figure 4(d), the overtaking lane is unavailable because of maintenance. Lanes decrease from two lanes to one lane. For cases in which a driving hazard and the vehicle come into contact, "collision" was displayed on the front screen, but running experiments continued.

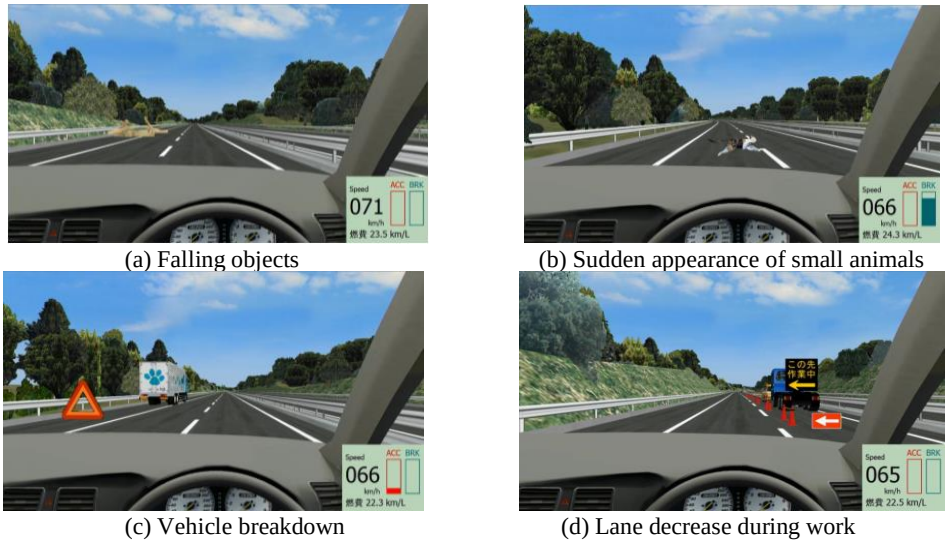
Next, we present an overview of running scenarios. To incorporate consideration of differences in visibility because of the time of day, we set the two conditions of daytime and nighttime, as well as the presence or absence of an overtaking vehicle as traveling conditions when encountering each driving hazard. Figure 5 presents driving scenarios of four types with control of the time of day (i.e., daytime or nighttime) and traffic flow (i.e., presence or absence of overtaking vehicles). The orders of occurrence of each driving topic were of two patterns: type A (i.e., falling objects → sudden appearance of small animals → breakdown vehicles → lane decrease) and type B (i.e., breakdown vehicles → sudden appearance of small animals → falling objects → lane decrease). The running time for one scenario is about 7–10 min. We gave instructions to each driver to observe the speed limit of 100 (km/h). Additionally, the maximum speed was limited automatically to 120 (km/h) on the DS side.

IV. Experiment Results and Arguments

Based on safety confirmation behaviors related to the driving hazards, we analyzed the reaction times of all drivers, elderly drivers, and young drivers at each measurement point. Finally, we assessed differences of driving characteristics between elderly drivers and young drivers, particularly addressing cognition, judgment, and operation. We will further develop the analysis of an earlier study [1] based on the influence of visibility because of changes in the driving time of day (daytime / nighttime) and changes in the perception of surroundings because of the presence or absence of overtaking vehicles. Specifically addressing the reaction time related to cognition, and that related to judgment and operation, we compare and analyze the respective driving characteristics of elderly drivers and young drivers.

A. Measurement points and driver response to driving hazards

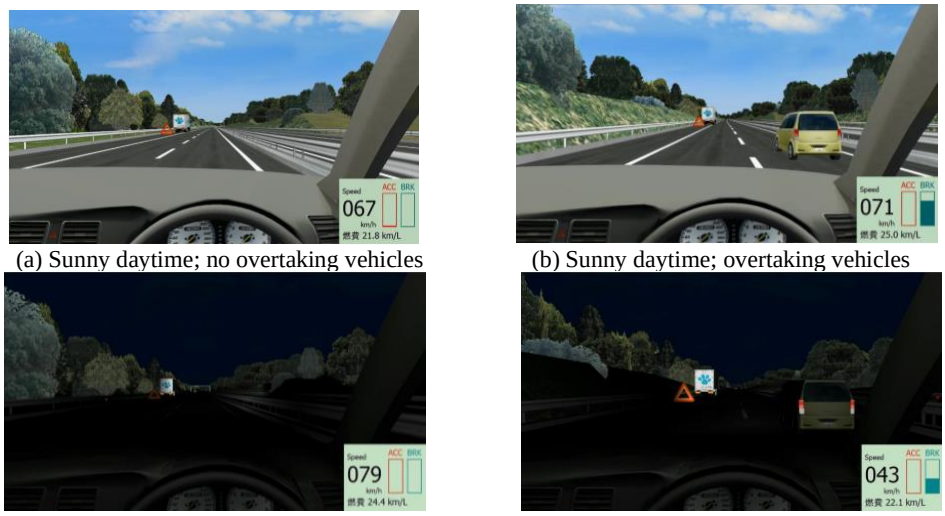
Specifically examining eye-gaze behaviors for these driving topics and safety confirmation behaviors related to lane change, we assess the correspondence between measurement points of driver reaction and eye-gaze information (i.e., heat maps, saccades) as portrayed in Figure 3. Figure 6 presents the time series changes of heat maps and saccades corresponding to each measurement point for the falling object. Each heat map represents the degree of gaze concentration in each of the following sections. They are from b) passing the reference point to c) falling object gazing, from c) falling object gazing to d) checking surrounding conditions, from d) checking surrounding conditions to e)



(a) Falling objects (b) Sudden appearance of small animals

(c) Vehicle breakdown (d) Lane decrease during work

Figure 4. Driving topics for running test.



(a) Sunny daytime; no overtaking vehicles (b) Sunny daytime; overtaking vehicles

(c) Sunny nighttime; no overtaking vehicles (d) Sunny nighttime; overtaking vehicles

Figure 5. Running scenarios.

avoidance actions. In addition, the time series changes of saccades are divided into two sections: from a) 5 s before passing the reference point to b) passing the reference point, and from b) passing from the reference point to e) avoidance action. Specifically examining the time series changes of heat maps and saccades in Figure 6, we can confirm driver states related to the attention to the falling object and the surrounding situation confirmation because of lane change. Consequently, based on the vehicle speed of b) the reference point during the passage, one can estimate the *reaction* time required for each action: the cognition, judgment, and operation for the falling object.

B. Driver Response to Falling Objects

Figure 7 presents the respective average reaction times of all drivers, elderly drivers, and young drivers at each measurement point for "falling objects" which is one driving topic. In Figure 7, the respective reaction times are shown.

1) from b) passing the reference point to c) falling object gazing,
2) from c) falling object gazing to d) surrounding situation confirmation,
3) from d) surrounding situation confirmation to e) avoidance behavior

Results for driving behaviors related to cognition, judgment, and operations for falling objects show that the elderly drivers tended to have longer reaction times than young drivers.

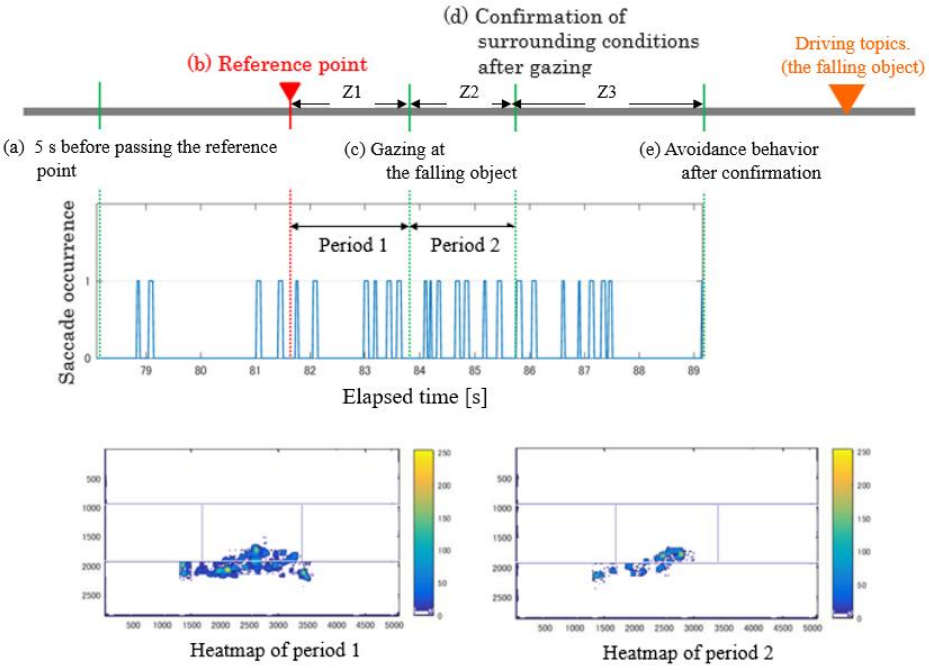


Figure 6. Time series changes of heatmaps and saccades corresponding to respective measurement points.

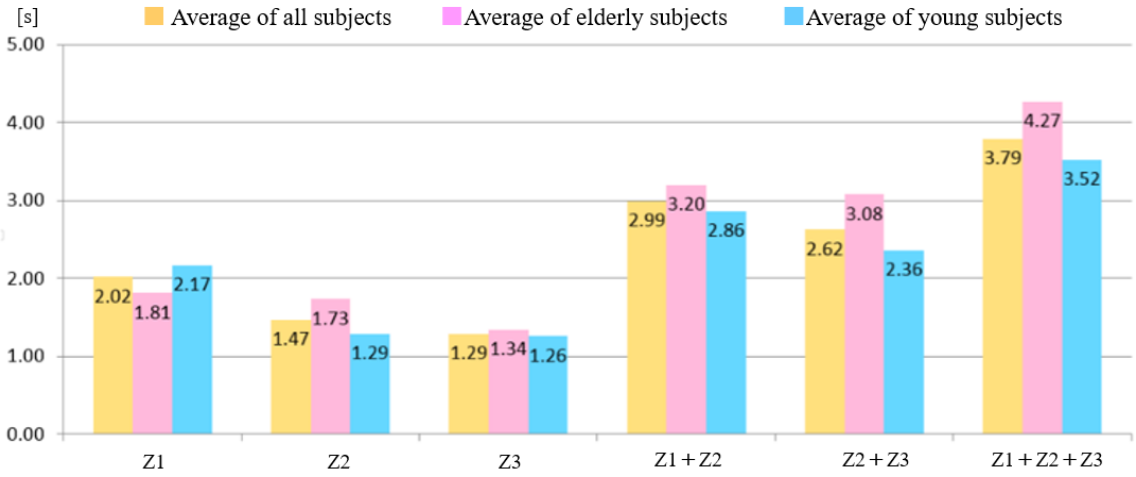
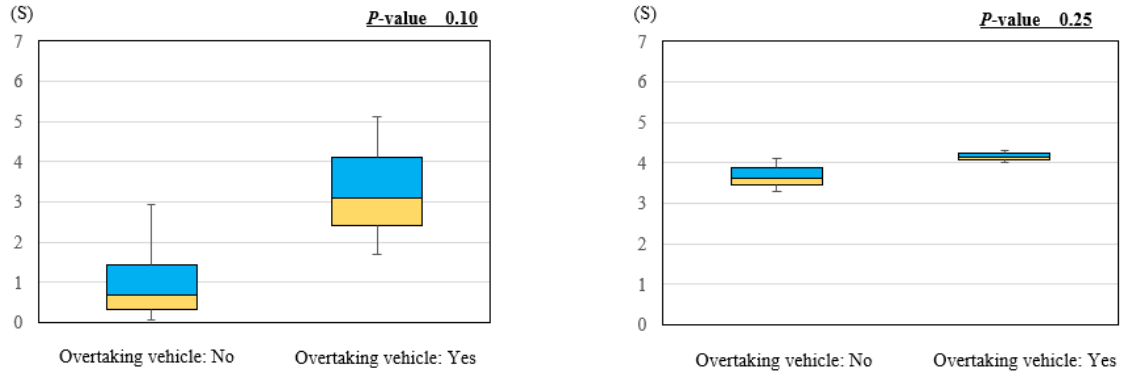


Figure 7. Driver reaction time to falling objects.

Furthermore, by classifying drivers based on the average of all participants and devoting attention to the relation with the driving style, we analyzed the existence of the following four groups.

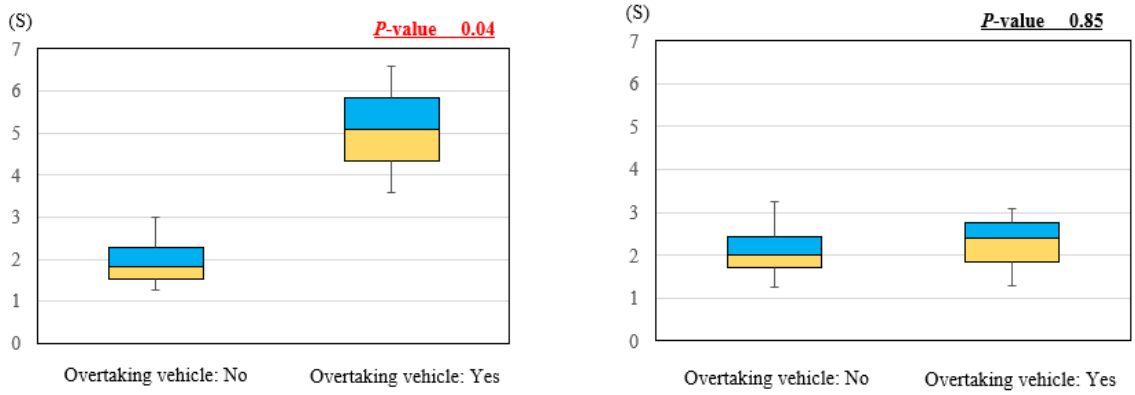
· Group A: Discovery and situation confirmation are quick; allowance for operation (i.e., cognition, judgment, and operation are all fast)
· Group B: Discovery is late, but situation confirmation and handling are quick (i.e., cognition is slow, but judgment and operation are fast)
· Group C: Discovery and situation confirmation are slow; operation is gradual (i.e., cognition, judgment, and operation are gradual)
· Group D: All reactions are average (i.e., cognition, judgment, and operation are all average)

Young drivers were classified into Group A or Group D, whereas the elderly drivers were more likely to be classified as Group B or Group C. We infer that driving behavior characteristics shift from group B to group C as drivers age.



(a) Young drivers (b) Elderly drivers

Figure 8. Response time for "cognitions" during daytime.



(a) Young drivers (b) Elderly drivers

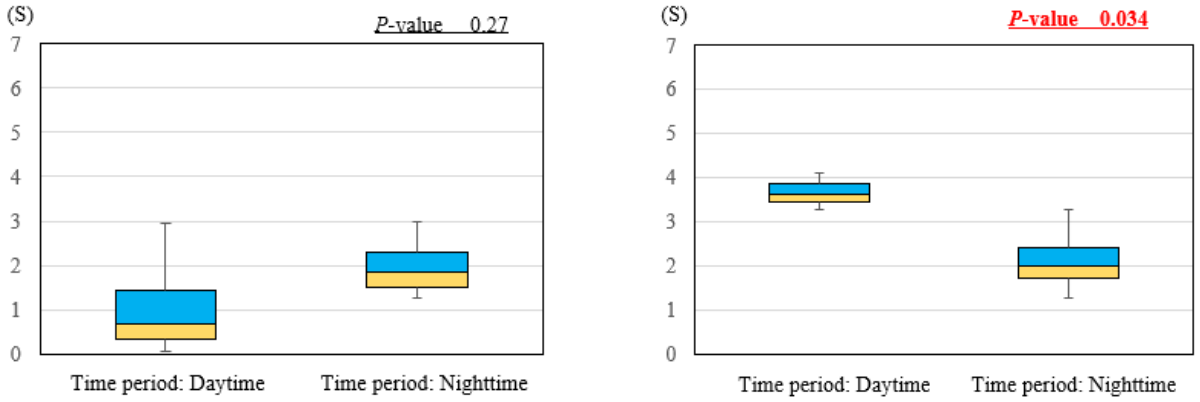
Figure 9. Response time for "cognitions" during nighttime.

C. Driver Response to Cognition

We next analyze driver reaction times for "cognition" from the viewpoint of surrounding recognition because of the presence or absence of overtaking vehicles. Figure 8 presents the reaction time of "cognition" during daytime. Figure 9 depicts the reaction time of "cognition" during nighttime. In these figures, panel (a) shows results for young drivers; panel (b) shows result for elderly drivers, clarifying differences of reaction time related to the presence or absence of overtaking vehicles. Particularly for young drivers in Figure 8 and Figure 9, the reaction time of "Overtaking vehicle: Yes" is longer than that of "Overtaking vehicle: No" for both daytime and nighttime. Especially, in "Time period: Nighttime", a significant difference (p = 0.04) was found. As a general characteristic, we can confirm that the reaction time tends to be longer at night with poor visibility. However, for elderly drivers, differences of the reaction time depending on the presence or absence of the overtaking vehicle are only slightly noticed. The reaction time of nighttime tends to be shorter than that of daytime.

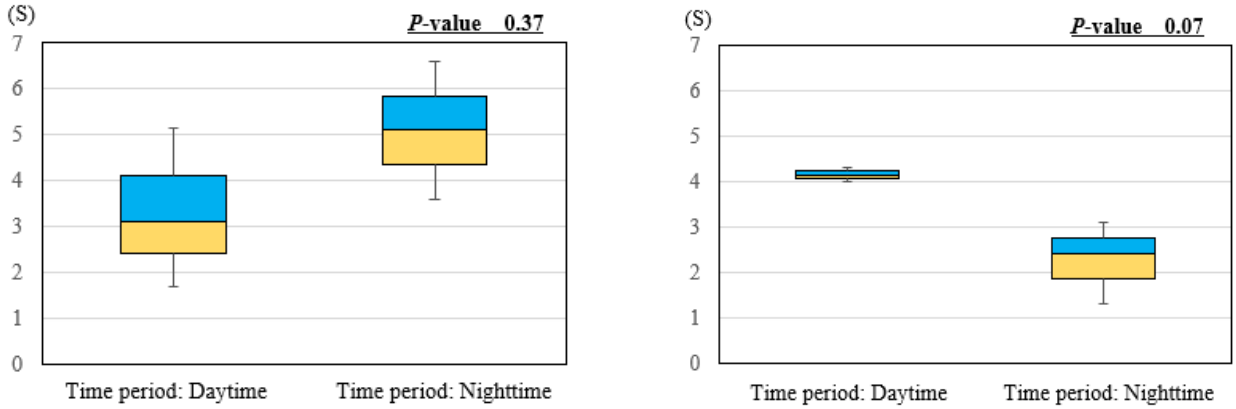
Next, we attempt to analyze visibility because of differences of the driving time of day (i.e., daytime and nighttime). Figure 10 presents the reaction time of "cognition" for "Overtaking vehicle: No". Figure 11 shows the reaction time of "cognition" for "Overtaking vehicle: Yes". In these figures, panel (a) shows results for young drivers; panel (b) presents results for elderly drivers. Specifically examining the elderly drivers in Figures 10 and 11, the response time of "Time period: Nighttime" is shorter than that of "Time period: Daytime", irrespective of the presence or absence of overtaking vehicles. A significant difference (p = 0.034) was found. By contrast, for young drivers, the difference of the reaction time between daytime and nighttime is slight, but reaction times of "Overtaking vehicle: Yes" tend to be longer than those of "Overtaking vehicle: No"

Based on the results described above, after summarizing conditions related to daytime and nighttime, and the presence or absence of overtaking vehicles, we analyze the reaction times and the moving distances required for "cognition" while comparing the elderly drivers to the young drivers.



(a) Young drivers (b) Elderly drivers

Figure 10. Response time for "cognitions" in "Overtaking vehicle: No".



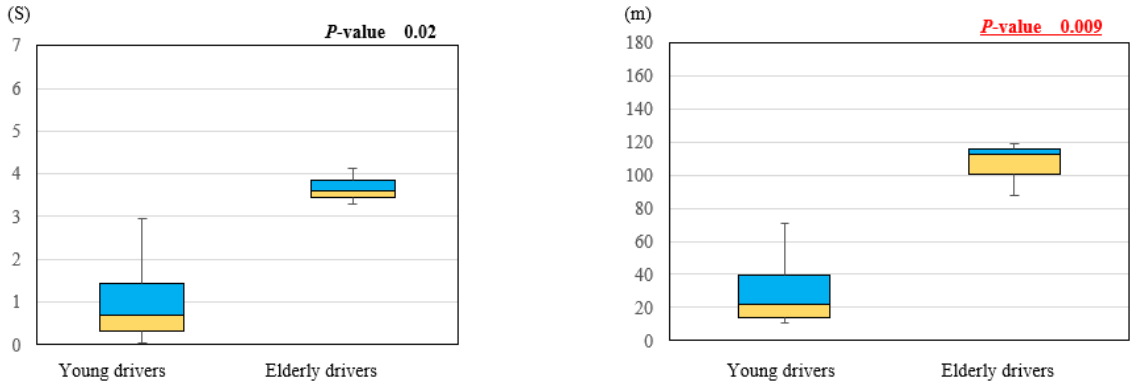
(a) Young drivers (b) Elderly drivers

Figure 11. Response time for "cognitions" in "Overtaking vehicle: Yes".

Figure 12 presents results obtained for "Time period: Daytime" and "Overtaking vehicle: No". Figure 13 presents results for "Time period: Nighttime" and "Overtaking vehicle: Yes". In these figures, panel (a) shows the reaction times of "cognition"; panel (b) shows the moving distance necessary for recognition, indicating the difference between elderly drivers and young drivers. By analyzing the results depicted in Figure 12, we can confirm that the reaction time and the moving distance of elderly drivers are longer than those of the young drivers. Especially, a significant difference (p = 0.09) in the moving distance was found. Actually, the tendency is reversed for results of Figure 13. The reaction time and the moving distance of the elderly drivers are shorter than those of the young drivers. Specifically examining the young drivers in Figures 12 and 13, one can confirm that the reaction time and the moving distance during nighttime are longer than those during daytime. Generally, visibility during nighttime is regarded as worse. Therefore, the results of young drivers are valid. However, the results obtained for elderly drivers in Figures 12 and 13 did not reflect the same tendency. The reaction time and the moving distance during nighttime were shorter than those during daytime. According to the questionnaire survey after the running test, the majority accounted for the elderly drivers as "the nighttime course made it easier to recognize falling objects". The reason for this trend might be that eye-gaze targets were limited during nighttime driving.

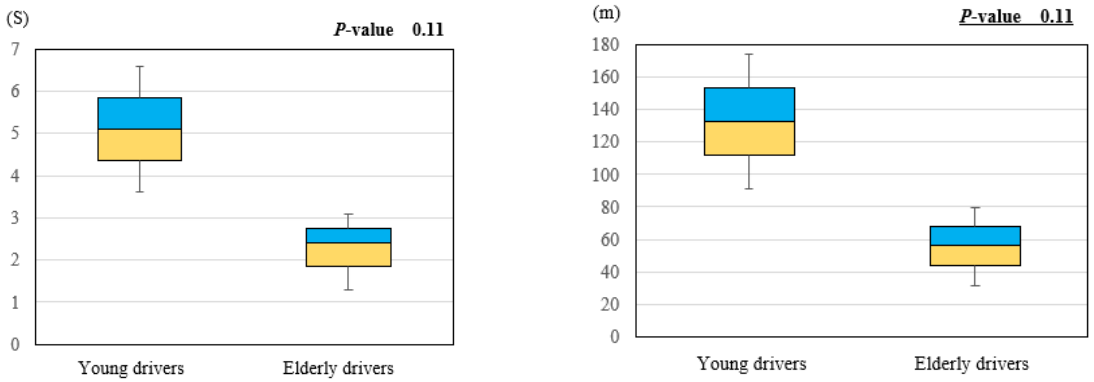
D. Driver Response to Judgment and Operation

We attempt to analyze the driver reaction time for "judgment and operation" from the viewpoint of recognition of the surroundings because of the presence or absence of overtaking vehicles. Figure 14 presents the reaction time of "judgment and operation" for daytime. Figure 15 portrays the reaction time of "judgment and operation" during nighttime. In these figures, panel (a) shows the results obtained for young drivers; panel (b) shows result obtained for elderly drivers, clarifying difference of reaction times related to the presence or absence of overtaking vehicles. Specifically examining the young drivers in Figures 14 and 15,



(a) Response time for "cognitions" (b) Moved distance for "cognitions"

Figure 12. Response time and moved distance for "cognitions" in "Time period: Daytime" and "Overtaking vehicle: No".



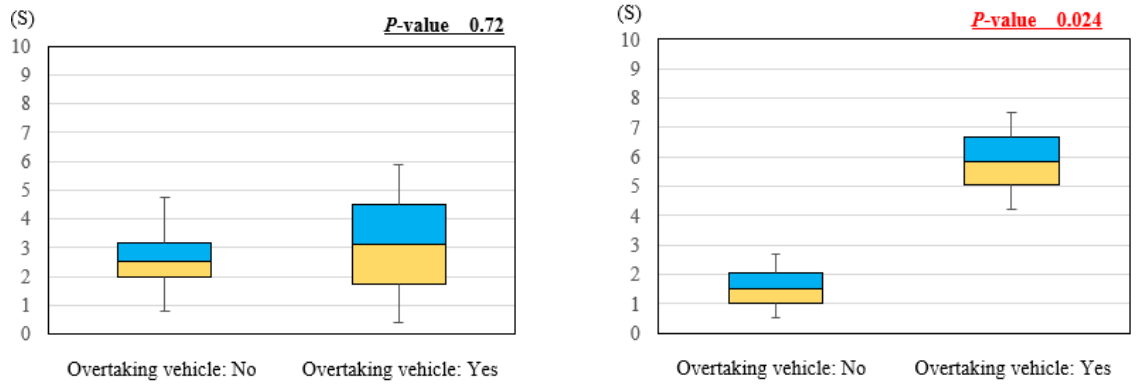
(a) Response time for "cognitions" (b) Moved distance for "cognitions"

Figure 13. Response time and moved distance for "cognitions" in "Time period: Nighttime" and "Overtaking vehicle: Yes".

differences of the reaction time depending on the presence or absence of the overtaking vehicle is noticed only slightly during daytime. However, the reaction time of "Overtaking vehicle: Yes" is longer than that of "Overtaking vehicle: No" during nighttime. A significant difference (p = 0.007) was found. However, for elderly drivers, the reaction time of "Overtaking vehicle: Yes" is longer than that of "Overtaking vehicle: No" during daytime. A significant difference (p = 0.024) was found. However, the difference of the reaction time related to the presence or absence of the overtaking vehicle is noticed only slightly at nighttime. Additionally, we can confirm that the reaction time of daytime tends to be shorter than that of nighttime.

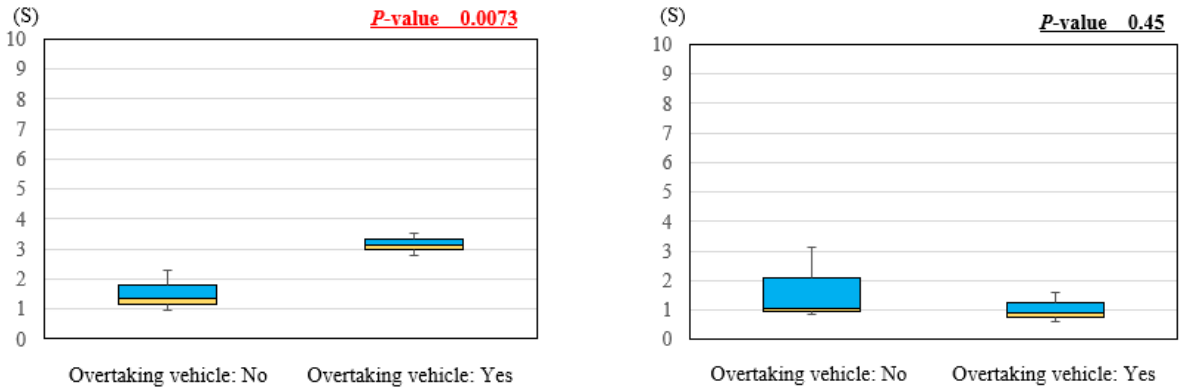
Next, we particularly address the environment in which the driving conditions are the most severe for each driver (i.e., "Time period: Nighttime" and "Overtaking vehicle: Yes"). Thereby, we attempt to analyze the response time for "judgment and operation" and their necessary moving distance by comparing elderly drivers to young drivers. Figure 16 presents the result for "Time period: Nighttime" and "Overtaking vehicle: Yes". In the figure, panel (a) shows the response time of "judgment and operation"; panel (b) shows the moving distance required for them. It shows the difference between elderly drivers and young drivers. The response time and moving distance of the elderly drivers are shorter than those of the young drivers. Particularly for the moving distance, a significant difference (p = 0.012) was found. The elderly drivers have more driving experience than the young drivers. We infer that the difference in driving behaviors related to "judgment and operation" after recognition resulted from their differences in driving experience.

Finally, to evaluate the state of peripheral recognition for each driver quantitatively, we specifically examine the number of saccades caused by differences in the driving time of day (i.e., daytime or nighttime), and compare data obtained for the elderly drivers and the young drivers. Figure 17 presents the number of saccades occurring per second related to "judgment and operation". In the figure, panel (a) shows the number of occurrences of "Time period: Daytime"; panel



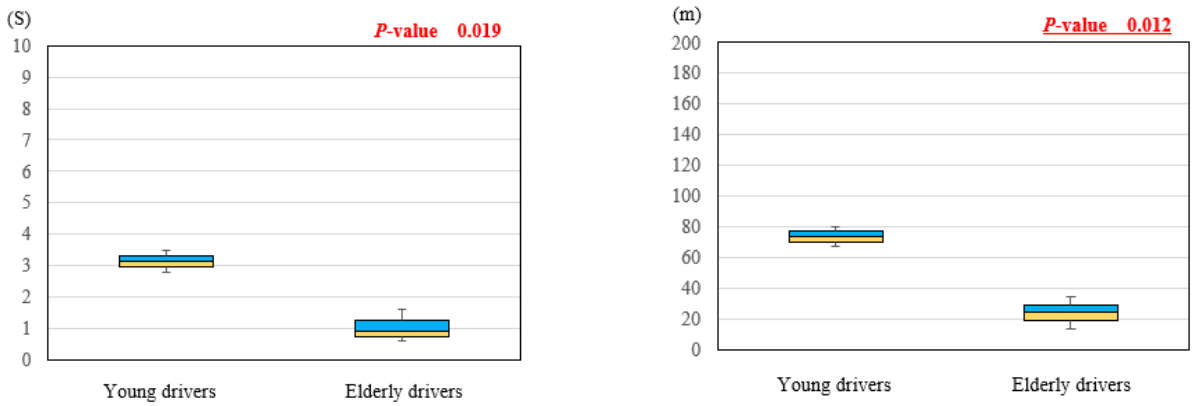
(a) Young drivers (b) Elderly drivers

Figure 14. Response time for "judgement and operations" during daytime.



(a) Young drivers (b) Elderly drivers

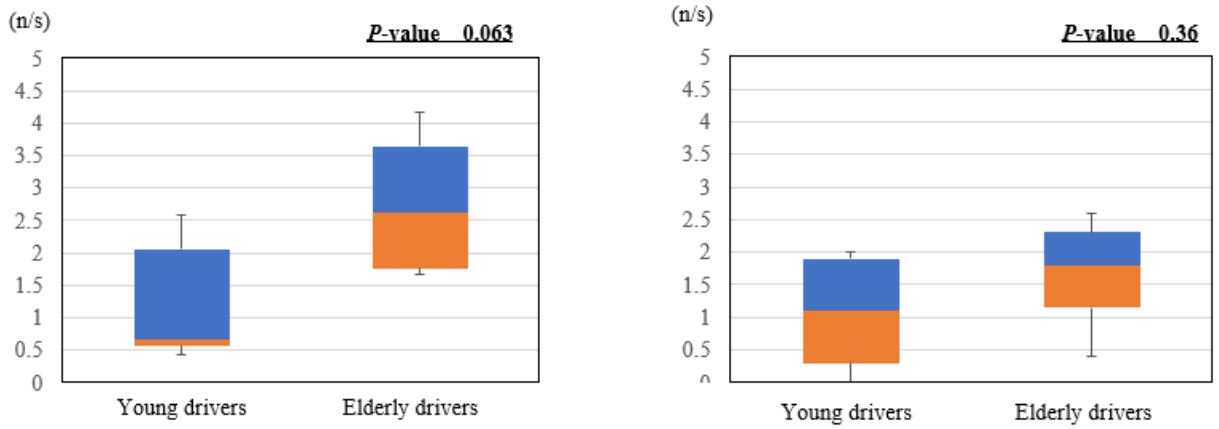
Figure 15. Response time for "judgement and operations" during nighttime.



(a) Response time for "judgement and operations" (b) Moved distance for "judgement and operations"

Figure 16. Response time and moved distance for " judgement and operations " in "Time period: Nighttime" and "Overtaking vehicle: Yes".

(b) shows the number of occurrences of "Time period: Nighttime". Differences between the elderly drivers and the young drivers are apparent. However, both (a) and (b) include datasets on the presence or absence of the overtaking vehicle. In "Time period: Daytime", the saccade frequency of the elderly drivers is higher than that of the young drivers. In contrast, their difference is not recognized in "Time period: Nighttime". The reason is the following. During the daytime



(a) Time period: Daytime (b) Time period: Nighttime

Figure 17. Number of saccade occurrences for "judgement and operations"

with good visibility, we surmise that elderly drivers performed more peripheral recognition for safety confirmation than young drivers did, based on their experience. During nighttime with poor visibility, the number of occurrences is suppressed by the effective field narrowing of each driver. The influence is inferred to have strongly affected the elderly drivers.

Taken together, the results described above confirmed that the reaction time related to "cognition" for these driving behaviors increases concomitantly with aging. In subsequent driving behaviors related to "judgment and operation", the elderly drivers tended to drive well based on their experience. However, in "Time period: Nighttime" with poor visibility, we assume that peripheral recognition for safety confirmation was disturbed by field narrowing of the elderly drivers.

V. Conclusion and Future Work

For this study, we defined four situations that are likely to trigger accidents on highways as driving hazards, i.e., vehicle breakdown, sudden appearance of small animals, falling objects, and lane decrease. By running experiments conducted while controlling the time period (daytime or nighttime) and the traffic flow (overtaking vehicle: Yes or No), we undertook comparative analysis for each driving hazard, particularly addressing driving characteristics of young drivers and elderly drivers related to "cognition", "judgment", and "operation". Results clarified the following points.

- Because of "cognition", the response time of young drivers of "Overtaking vehicle: Yes" was longer than that of "Overtaking vehicle: No" in both daytime and nighttime. Particularly, in "Time period: Nighttime", significant difference (p = 0.04) was found.
- Regarding elderly drivers, irrespective of the presence or absence of an overtaking vehicle, the response time related to "cognition" of "Time period: Nighttime" was shorter than that of "Time period: Daytime". Particularly, for "Overtaking vehicle: No", significant difference (p = 0.034) was found.
- For "Time period: Daytime" and "Overtaking vehicle: No", the response time and the moved distance related to "cognition" of the elderly drivers are longer than that of the young drivers. Particularly, significant difference of the moved distance (p = 0.09) was found.
- Regarding young drivers, the response time because of "judgment and operations" of "Overtaking vehicle: Yes" was longer than that of "Overtaking vehicle: No" in nighttime. Significant difference (p = 0.007) was found.
- Regarding elderly drivers, the response time because of "judgment and operations" of "Overtaking vehicle: Yes" was longer than that of "Overtaking vehicle: No" in daytime. Significant difference (p = 0.024) was found.
- For "Time period: Nighttime" and "Overtaking vehicle: Yes", the response time and the moved distance related to "cognition" of the elderly drivers was shorter than that of the young drivers. Particularly, significant difference of the moved distance (p = 0.012) was found.

Future work will be undertaken to construct a dangerous-driving prediction model based on findings obtained from this study with the aim of improving the prediction model accuracy.

ACKNOWLEDGMENT

This work was supported by a technical research grant of East Nippon Expressway Company Ltd.

References

[1] K. Sato, Y. Oomomo, H. Madokoro, M. Ito, and S. Kadowaki, "Analysis of Cognitive Functions in the Encountered State of Driving Topics on Expressways," Proceedings of the Fourteenth International Multi-Conference on Computing in the Global Information Technology (ICCGI2019), pp.1-8, July 2019.

[2] K. Kamiji and N. Takahashi, "Achievement and Effort of Safety Traffic Society for Elderly Drivers by JAMA," International

Association of Traffic and Safety Sciences, vol. 35, no. 3, pp. 221-227, 2011.

[3] H. Suzuki, "Motivating Senior Drivers toward Traffic Safety: The Traffic Sociology Viewpoint," International Association of Traffic and Safety Sciences, vol. 35, no. 3, pp. 194-202, 2011.

[4] Y. Nakano, T. Kojima, H. Kawanaka, and K. Oguri, "Improvement of Elderly Driver's Ability by Pointing and Calling Method," The Institute of Electronics, Information and Communication Engineers D, vol. J97-D, no.1, pp. 135-144, 2014.

[5] T. Tanaka, K. Fujikake, T. Yonekawa, M. Yamagishi, M. Inagami, F. Kinoshita, H. Aoki, and H. Kanamori, "Analysis of Relationship between Forms of Driving Support Agent and Gaze Behavior: Study on Driver-Agent for Encouraging Safe driving Behavior of Elderly Drivers," IEICE Technical Report, 117(72), pp. 13-18, 2017.

[6] H. Aoki, H. Kanamori, M. Yamagishi, T. Tanaka, and T. Yonekawa, "Study on Driver Characteristics for Delaying Driving Cessation (1), – Construction of Human, Aging, and Driving Characteristic Database for Elderly Drivers –," Society of Automotive Engineers of Japan, Inc. (JSAE), Annual Conference (Spring) Proceedings, 2015.

[7] T. Tanaka, K. Fujikake, T. Yonekawa, M. Yamagishi, M. Inagami, F. Kinoshita, H. Aoki, and H. Kanamori, "Driver Agent for Promoting Driving Behavior Improvement of Elderly," Human-Agent Interaction Symposium 2016, G-4, 2016.

[8] T. Tanaka, T. Yonekawa, H. Aoki, M. Yamagishi, M. Inagami, I. Takahashi, and H. Kanamori, "Analysis of Relationship between Driving Behavior and Biofunction of Drivers including Elderly at Intersection with a Stop Sign: – Study on Driver Characteristics for Delaying Driving Cessation –," Society of Automotive Engineers of Japan, Inc. (JSAE), Vol. 48, No. 1, pp.147-154, 2017.

[9] T. Tanaka, K. Fujikake, T. Yonekawa, M. Yamagishi, M. Inagami, F. Kinoshita, H. Aoki, and H. Kanamori, "Study on Driver Agent based on Analysis of Driving Instruction Data – Driver Agent for Encouraging Safe Driving Behavior (1) –," IEICE Transactions on Information and Systems, Vol. E101-D, No. 5, 2018.

[10] K. Fujikawa, T. Tanaka, T. Yonekawa, M. Yamagishi, M. Inagami, F. Kinoshita, H. Aoki, and H. Kanamori, "Comparison of Subjective Evaluation of Different Forms of Driving Agents by Elderly People," Japan Human Factors and Ergonomics Society, Vol.53, pp.214-224, 2017.

[11] K. Fujikake, T. Tanaka, Y. Yoshihara, T. Yonekawa, M. Inagami, H. Aoki, and H. Kanamori, "Effect of Driving Behavior Improvement by Driving-Support and Feedback-Support of Driver-Agent," Society of Automotive Engineers of Japan, Inc. (JSAE), Vol. 50, No. 1, pp. 134-141, 2019.

[12] M. Takahara, M. Kokubun, T. Wada, and S. Doi, "Effects of a Stopping-situation Assistance System for Elderly Drivers," International Association of Traffic and Safety Sciences, Vol. 36, No. 1, pp. 6-13, 2011.

[13] K. Iida, "Hypothesis Construction on Reverse Running Occurrence Process on Highway," Grant Research by Takata Foundation ISSN2185-8950, Available from URL: http://www.takatafound.or.jp/support/articles/pdf/150626_10.pdf [retrieved: May 2020]

[14] G. Abe, K. Kikuchi, R. Iwaki, and T. Fujii, "Effects of Cognitive Distraction on Driver's Visual Attention," The Society of Mechanical Engineers (JSME), Transactions of the JSME (c), Vol. 76, No. 767, pp. 1662-1668, 2010.

[15] R. Honma, G. Abe, and K. Kikuchi, "Characteristics of Visual Attention while Driving under the State of Drowsiness," Society of Automotive Engineers of Japan, Inc. (JSAE), Vol. 42, No. 5, pp. 1217-1222, 2011.

[16] Council for Science, Technology and Innovation in Cabinet Office: Cross-ministerial Strategic Innovation Promotion Program (SIP) Innovation of Automated Driving for Universal Services, pp.2, 2017.

[17] S. Wada, T. Hagiwara, H. Hamaoka, Y. Ninomiya, M. Tada, and T. Ohiro, "Comparative Study on Situation Awareness between Elderly and General Drivers on the Expressway using Highly Automated Vehicle," Society of Automotive Engineers of Japan, Inc. (JSAE), Technical Paper, 2018.

[18] Society of Automotive Engineers (SAE): Autonomous car driving levels (SAE J3016), Sep. 2016.

[19] C. Miyajima, S. Yamazaki, T. Bando, K. Hitomi, H. Terai, H. Okuda, T. Hirayama, M. Egawa, T. Suzuki, and K. Takeda, "Analyzing Driver Gaze Behavior and Consistency of Decision Making During Automated Driving," 2015 IEEE Intelligent Vehicles Symposium (Ⅳ), Seoul, Korea, 2015.

[20] L. J. Molnar, A. K. Pradhan, D. W. Eby, L. H. Ryan, R. M. St. Louis, J. Zakrajsek, B. Ross, B. Lin, C. Liang, B. Zalewski, and L. Zhang, "Age-Related Differences in Driver Behavior Associated with Automated Vehicles and Transfer of Control between Automated and Manual Control: A Simulator Evaluation," Technical Report, UMTRI 2017-4, University of Michigan Transportation Research Institute, 2017.

[21] M. Ishibashi, M. Okuwa, S. Doi, and M. Akamatsu, "HQL Driving Style Questionnaire: DSQ," Research Institute of Human Engineering for Quality Life, 2003.

[22] M. Ishibashi, M. Okuwa, S. Doi, and M. Akamatsu, "HQL Workload Sensitivity Questionnaire: WSQ," Research Institute of Human Engineering for Quality Life, 2003.

Topological Reduction of Stationary Network Problems: Example of Gas Transport

Anton Baldin[(1)], Tanja Clees[(1,2)], Bernhard Klaassen[(1)],
Igor Nikitin[(1)], Lialia Nikitina[(1)]

[(1)] Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany
[(2)] University of Applied Sciences Bonn-Rhein-Sieg, Sankt Augustin, Germany
Email: Name.Surname@scai.fraunhofer.de

***Abstract*—The general method of topological reduction for the network problems is presented on example of gas transport networks. The method is based on a contraction of series, parallel and tree-like subgraphs for the element equations of quadratic, power law and general monotone dependencies. The method allows to reduce significantly the complexity of the graph and to accelerate the solution procedure for stationary network problems. The method has been tested on a large set of realistic network scenarios. Possible extensions of the method have been described, including triangulated element equations, continuation of the equations at infinity, providing uniqueness of solution, a choice of Newtonian stabilizer for nearly degenerated systems. The method is applicable for various sectors in the field of energetics, including gas networks, water networks, electric networks, as well as for coupling of different sectors.**

Keywords–modeling of complex systems; topological reduction; globally convergent solvers; applications; gas transport networks.

I. Introduction

This work is an extension of our conference paper [1], where the method of topological reduction for stationary gas transport network problems has been introduced. We have performed additional testing of the method on a large number of networks, evaluated various statistical characteristics, improved main algorithms for representing the equations and for their solution, considered the extension of the method for the networks of general type, as well as their coupling.

The physical modeling of gas transport networks is comprehensively described in works [2]–[4]. The element equations for pipes vary from the simplest quadratic form to more complex formulae by Nikuradze, Hofer and Colebrook-White. In our papers [5][6], we have shown how to continue these formulae to the whole domain of model variables, in order to achieve a global convergence for non-linear solvers. Further, in paper [7] we have constructed a universal translation algorithm, capable of formulating network problems for non-linear solvers with arbitrary problem description language. In paper [8], we presented theoretical foundations of topological reduction methods for generic stationary network problems.

In this paper, we continue the development of general topological reduction methods, applied to gas transport networks as an example. Our motivation is to accelerate solution procedure for stationary network problems. The goal is to perform significant reduction of the graphs, preserving the accuracy of the modeling. The main idea is to reduce the series and parallel connections of elements in the network, with the operations, known in the theory of Series-Parallel Graphs (SPG) [9]. These operations can also be extended by contraction of a leaf, which after recurrent application contracts tree-like subgraphs, leading to Generalized Series-Parallel Graphs (GSPG) [10]. Such elementary operations are shown in Figure 1. In paper [8], we have estimated the efficiency of this method and shown on realistic gas transport networks that high reduction factors can be achieved. In our current work, we perform an actual implementation of the topological reduction for pipes, which form a considerable part of the gas transport networks.

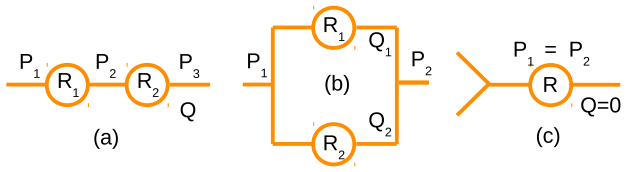


Figure 1. Main operations in GSPG reduction: series (a), parallel (b) connections to be reduced, contraction of a leaf (c).

In Section II, we present the details of a topological reduction procedure for pipes, modeled by quadratic friction law. In Section III, the results of numerical experiments with estimation of reduction factors and acceleration rates are given. In Section IV, we perform a comparison of our method with [11], which is also based on graph theory but using a different approach. In Section IV, we also discuss possible extensions of our method. In Section V, we present further improvement of the main algorithms.

The described algorithms are implemented in the software MYNTS (Multi-phYsics NeTwork Simulator) [12], developed in our group.

II. Topological Reduction Algorithm for Pipe Networks

For the equations representing the pipes, one can use the simplest quadratic friction law from [2][11]:

$$P_{in}|P_{in}| - P_{out}|P_{out}| = RQ|Q|, \qquad (1)$$

where $P_{in,out}$ are the input and output pressures and Q is the mass flow through the pipe. R is a resistance coefficient, depending on the pipe length L, diameter D, roughness parameter k, universal gas constant R_{gas}, temperature T, compression factor z and molar mass μ:

$$R = 16L/(\pi^2 D^5)/(2\log_{10}(D/k) + 1.138)^2 \times R_{gas} T z/\mu \cdot 10^{-10}. \qquad (2)$$

All parameters are given in SI units (French, Système International), except of pressures, given in bar, hence the scale factor at the end of the formula. The structure of the term $Q|Q|$ ensures the symmetry of the equation when reversing the flow direction $Q \to -Q$. The similar structure of P-terms has a very special reason: it provides a monotonic continuation of the equation to the non-physical domain $P < 0$.

It was shown in [6] that, as a result of such continuation, the solver maintains stability also in the non-physical domain, where it can occasionally wander during the iterations. In addition, with such an extension (and the refinements, done here in Section V), the system describing the stationary state of the network has a unique solution, even if the problem was set infeasibly. The simplest example of such an infeasible setting is to take a real network, such as shown in Figure 2, require a large throughput from suppliers to consumers, but at the same time switch off all the compressors. This problem, obviously, will not have a solution. On the other hand, if one uses the techniques from [6], the solution will exist and will be unique even in this case, but it will be located in the nonphysical domain $P < 0$. Thus, in this approach, one has a necessary and sufficient *feasibility indicator*, lacking for other solvers, for which the infeasible statement of the problem is indistinguishable from the occasional divergence.

Let us consider the above described GSPG elementary operations for pipe networks.

The series connection is (see Figure 1a):

$$\begin{aligned} P_1|P_1| - P_2|P_2| &= R_1 Q|Q|, \\ P_2|P_2| - P_3|P_3| &= R_2 Q|Q|. \end{aligned} \tag{3}$$

From here, we add the 2 formulas to get:

$$P_1|P_1| - P_3|P_3| = R^s_{12} Q|Q|, \;\; R^s_{12} = R_1 + R_2. \tag{4}$$

The inverse reconstruction of the eliminated variable P_2 is:

$$P_2|P_2| = P_1|P_1| - R_1 Q|Q|. \tag{5}$$

The parallel connection is (see Figure 1b):

$$\begin{aligned} P_1|P_1| - P_2|P_2| &= R_1 Q_1|Q_1| = R_2 Q_2|Q_2|, \\ Q &= Q_1 + Q_2. \end{aligned} \tag{6}$$

From here, we solve this system for $Q_{1,2}$ to get:

$$\begin{aligned} P_1|P_1| - P_2|P_2| &= R^p_{12} Q|Q|, \\ R^p_{12} &= \left(R_1^{-1/2} + R_2^{-1/2} \right)^{-2}. \end{aligned} \tag{7}$$

The inverse reconstruction of the eliminated variables $Q_{1,2}$ is:

$$\begin{aligned} Q_1 &= Q/((R_1/R_2)^{1/2} + 1), \\ Q_2 &= Q/((R_2/R_1)^{1/2} + 1). \end{aligned} \tag{8}$$

Contracting the leaf, see Figure 1c, in the simplest case of zero flow results in the removal of P_2, Q variables. The inverse reconstruction consists of the setting $Q = 0$ and copying $P_2 = P_1$.

It should be noted that there are two types of source/sink nodes in gas networks. Q_{set} is the node in which the flow is set. P_{set} is the node where the flow is not fixed, but the pressure is set. For parallel connections, nodes of this type at the ends do not pose a problem. For series connections, the presence of such specifiers in the intermediate node leads to deviations from Kirchhoff's law and represents an obstacle to the reduction. Next, we discuss a special algorithm that allows to move the Q_{set} specifiers over the network. In combination with it, the reduction can be continued.

For contraction of the leaf, the P_{set} specifier represents an obstacle, because when shifting to the neighboring node, the P_{set} specifier gets an unfixed pressure value that depends on the flow. To contract a leaf with the Q_{set} specifier, two options are possible. First, block contracting leafs with a nonzero Q_{set}. As a result, the reduction will be incomplete, but the end Q_{set} nodes will be intact, which is convenient for formulating scenarios with different values of Q_{set} and for controlling the feasibility condition $P > 0$ at endpoints. Second, allow such leafs to be moved, with Q_{set} moving to the other side and summing it up with another Q_{set} that may be located there. For the inverse reconstruction, the value of Q_{set} must be saved, after that the inverse operations can be performed. The pressure at the free end is not determined by simple copying, but is found from the equation of the element:

$$P_2|P_2| = P_1|P_1| - R Q_{set}|Q_{set}|. \tag{9}$$

III. The results

We have implemented GSPG reduction algorithm with fixed Qsets and tested it on three realistic networks. The simplest network N1 is shown in Figure 2. It includes 4 compressors (2 stations with 2 compressors each), 2 Psets (shown by rhombi n56, n99) and 3 Qsets (triangles n76, n80, n91). Originally (level0), the network contains N=100 nodes and E=111 edges, including P=34 pipes. Then (level1), a topological cleaning algorithm from [8] is used, removing (if any) parts of the graph, disconnected from pressure suppliers, as well as contracting superconducting edges, such as shortcuts, open valves and short pipes ($D = L = 1$ m). This operation is absolutely necessary for the stability of the solver, since disconnected parts possess undefined pressure and loops of superconducting edges have undefined circulating flow. This level of reduction looks similar to level0, just some valves, shortcuts and internals of stations are removed. The total count on this level is N=39, E=40, P=34.

After that (level2), GSPG reduction with fixed Qsets is applied, leaving N=13, E=14, P=8 elements. This corresponds to the reduction factor 2.9. Then, we have implemented all necessary GSPG operations described by the formulae above. For the solution procedure, after the reduction, we obtain the acceleration factor 2.2. The solution on level2 is identical with level1 up to the solver tolerance (set to tol=10^{-5} in our numerical experiments).

For GSPG reduction with moving Qsets (level3), we have implemented the formal reduction algorithm, sufficient for the estimation of the reduction factor. On this level, we have N=8, E=9, P=3 elements, comprising the reduction factor 1.6 relative to the previous level. The numerical counterpart of the algorithm has not been implemented yet, that is why the reduced network for level3 on Figure 2 does not have pressure data. In the next section, we will discuss the details of Qset movement algorithm necessary for this level.

The same tests have been performed on more complex networks N2 and N3, provided by our industrial partners for benchmarking. The parameters of the networks and the results

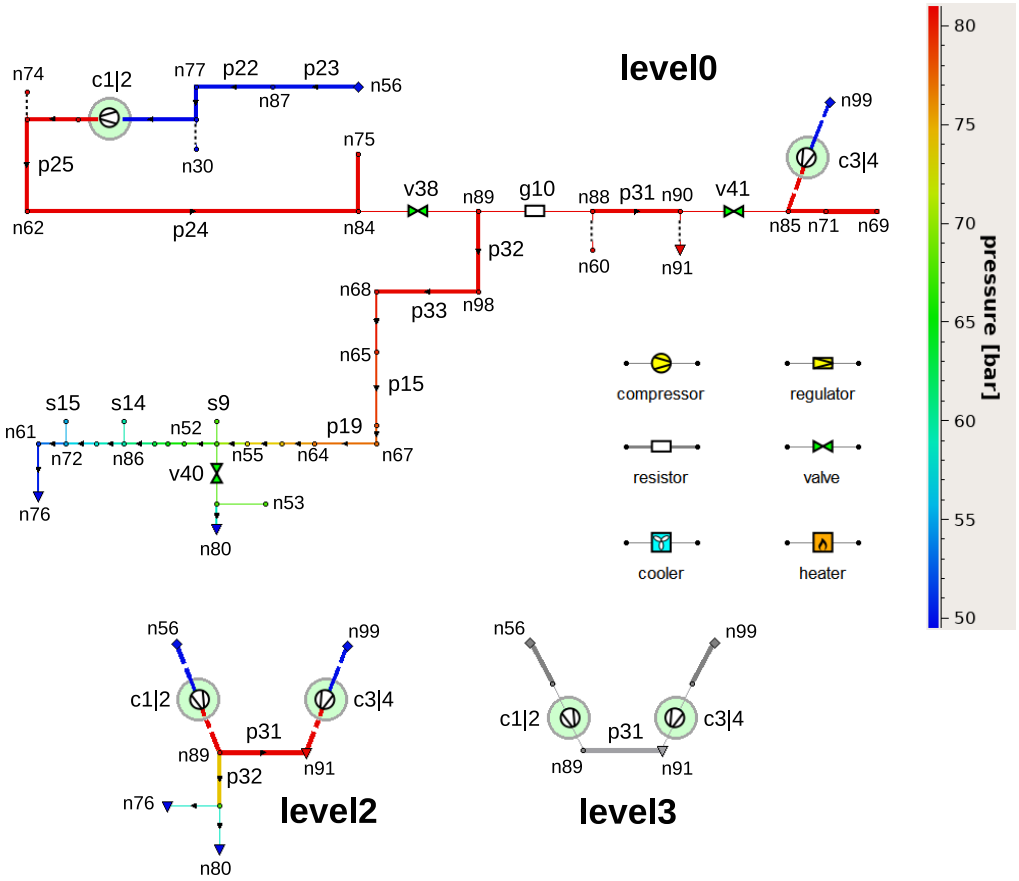


Figure 2. Realistic gas transport network N1 at different reduction levels: level0 = original network; level2 = GSPG reduction with fixed Qsets; level3 = GSPG reduction with moving Qsets. (Not shown: level1 = removing disconnected parts and superconductive elements, looking similar to level0.)

TABLE I. PARAMETERS OF TEST NETWORKS

network	compressors	regulators	Psets	Qsets
N1	4	0	2	3
N2	7	18	4	64
N3	25	54	6	290

TABLE II. NODES:EDGES:PIPES COUNT FOR DIFFERENT REDUCTION LEVELS

network	level0	level1	level2	level3
N1	100:111:34	39:40:34	13:14:8	8:9:3
N2	973:1047:500	528:541:479	198:208:146	126:134:72
N3	4721:5362:1749	1723:1814:1666	705:755:607	296:332:184

of the reduction are presented in Tables I-III. The obtained level1/level2 reduction factors vary in the range 2.4-2.9, while acceleration factors solve1/solve2 are 2.2-2.6. The 'filter' step in Table III includes the necessary preprocessing and reduction

TABLE III. TIMING FOR TWO REDUCTION LEVELS*

network	level1		level2	
	filter	solve	filter	solve
N1	0.006	0.044	0.009	0.02
N2	0.063	0.5	0.09	0.196
N3	0.243	2.103	0.371	0.944

* in seconds, for 3 GHz Intel i7 CPU 8 GB RAM workstation; 'filter' includes removing disconnected parts and superconductive elements (for level1,2) and GSPG reduction (for level2); 'solve' includes translation procedure, actual solving and extracting the result; the actual solving is performed with IPOPT.

of the networks. The 'solve' step includes translation of the network to the form suitable for the solver and the solution procedure itself, which share the timing in 1:1 proportion. Currently, our system uses the universal translation algorithm from [7]. It allows to plug in generic non-linear solvers with an arbitrary problem description language, requiring only to adjust a translation matrix in the algorithm. In particular, we have experimented with IPOPT (Interior Point OPTimizer) [13], Mathematica [14], MATLAB (MATrix LABoratory) [15] and a Newton solver, developed in our group. The best results for our type of problems have been obtained with IPOPT and Newton,

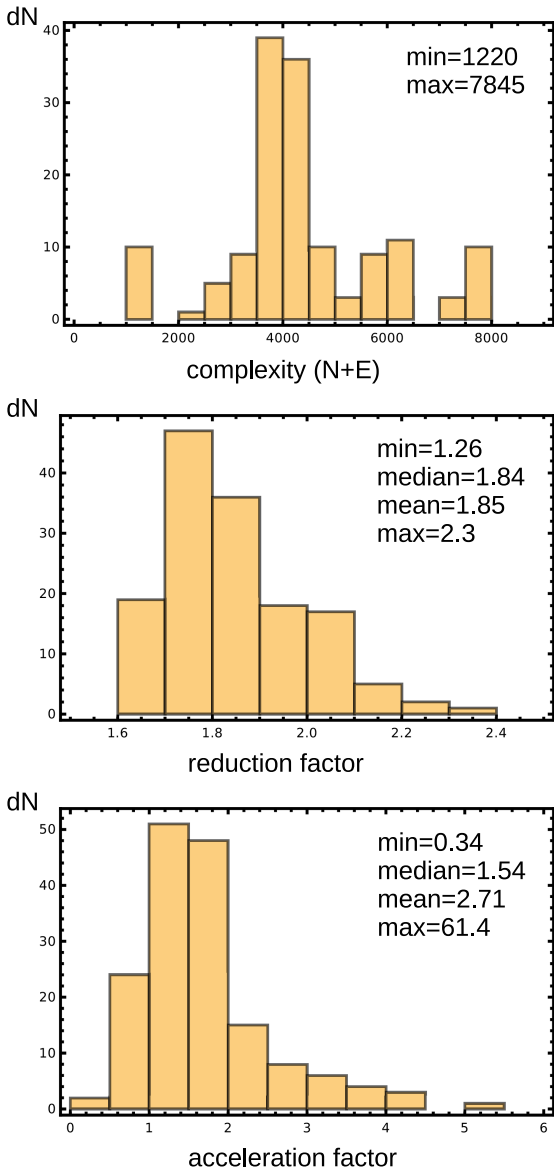


Figure 3. Topological reduction characteristics for additional 146 gas transport networks.

while these two solvers among themselves have comparable performance. The details of the implementation of the Newton solver will be published elsewhere.

The solution procedure involves a multiphase workflow, described in [7]. Although global convergence from an arbitrary starting point for stationary network problems is guaranteed theoretically [6], the multiphase procedure is still empirically faster. This procedure gradually increases the complexity of the modeling and uses the result of the previous phase as a starting point for the next one. In our numerical experiments, a 3-phase procedure is used, relevant to the modeling of compressors and regulators in the network. In the first phase, compressors and regulators have enforced goals, e.g., $P_{out} = Const$. Then, they are set to a simplified universal *free model* and, finally, to the individually calibrated *advanced model* [8]. The timing in Table III presents the sum over 3 phases.

More testing: in this paper, we performed the measurement of topological reduction characteristics for additional 146 gas transport networks of different complexity. The results are shown on Figure 3. The upper histogram shows the distribution of complexity value $N + E$ at the reduction level 1. This complexity value is more appropriate than the nominal one at level 0, since it counts only the active parts of the networks and only non-trivial (resistive) elements. The complexity values between approximately 1K and 8K elements are populated, with the maximum at 4K. The reduction factor between level 1 and level 2 is shown in the middle image. The reduction varies between 1.26 and 2.3, with median value 1.84. The shape of the histogram resembles Poisson distribution.

The acceleration factor is shown on the bottom image. It varies between 0.34 (deceleration) and 61.4 (strong acceleration), with median 1.54 and mean 2.71. The distribution is also looking like Poisson one, with outliers. The reason for these outliers is related with a randomness of the solver path towards the solution. Even with the regularization, the stationary problem formulation $f(x) = 0$ contains a complicated landscape of the function f, including sharp hills and narrow valleys, which lead to a slowdown of the solution procedure. The location of such features is random, so that topologically reduced problem can stuck in such regions, while the non-reduced problem was solved smoothly, leading to deceleration. Vice versa, the topological reduction can help to avoid these problematic regions, bringing very strong acceleration factor. Both cases are visible on the histograms as outliers.

Our statistics shows that deceleration events happen at a small probability and acceleration of solution prevails. On the other hand, strongly accelerated outliers skew the distribution, so that mean differs from the median significantly. While the median presents a statistical center of the distribution, mean value can be important in practically relevant setup, when a large number of the network cases is solved as a whole. Such applications appear, for example, in ensemble simulations, when stability of solution, sensitivity to variation of parameters or other statistically relevant characteristics are evaluated. The total time of ensemble simulation sums the individual time of each solution, making the mean time an important characteristic. Note that not the mean of the ratios between the original t_1 and reduced topology t_2 timings should be taken, but the ratio of corresponding means. In our testing cases, however, these characteristics are similar: $\langle t_1/t_2 \rangle = 2.71$, while $\langle t_1 \rangle / \langle t_2 \rangle = 2.69$.

IV. Possible extensions

At first, we perform a comparison with paper [11], where a different approach for topological reduction was taken. Then, we describe possible generalizations of our topological reduction algorithm.

a) Comparison with paper [11]: in this paper, the stationary problem in gas transport networks was studied, where subgraphs consisting of pipes only were considered. The pipes were modeled by the expressions of type (1) and the 2nd Kirchhoff law was consistently applied, by summing this expression over independent cycles in the subgraph. As a result, P-variables drop off from such sums and a system of smaller size depending only on Q-variables remains, for which the existence and uniqueness of the solution is proven.

Although the approach looks promising, for its practical implementation, some problems exist.

This approach does allow to reduce the dimension of the system by extracting from it a subsystem that depends only on Q-variables. The dimension of the subsystem is equal to the number of independent cycles in the subgraph. The subsystem has a unique solution for which, however, it is generally impossible to obtain an analytic expression. Thus, it should be solved numerically, for example, by Newton's method. The remaining variables in the subgraph are obtained by an unambiguous analytical reconstruction procedure. The problem appears when this subgraph is considered in the context of a complete graph containing other elements than pipes, for example, compressors. The solution of the complete problem is usually also found by the Newton's method. For the subgraph, this means that the solution must be found many times, with variable boundary conditions. In this case, a combination of two Newton's methods, external and internal, will require from the subgraph not only a solution, but also its derivatives with respect to the boundary conditions. Such a combination is in any case not an efficient way to solve the system.

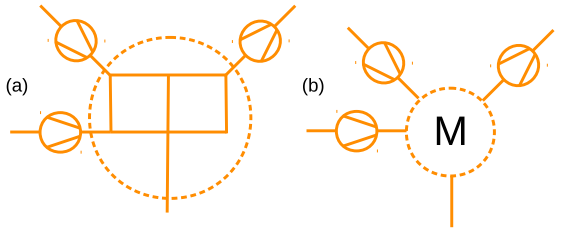


Figure 4. Shrinking a subgraph (a) creates a generalized network element with a fixed number of pins (b), a multipin (M).

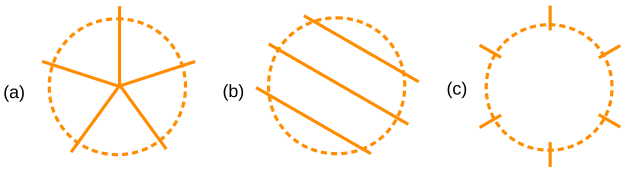


Figure 5. Particular examples: (a) 5-pin star; (b) 6-pin with 3 parallel connections; (c) empty 6-pin. In all cases, the number of equations describing the multipin is equal to the number of pins.

Another problem is that, according to [11], a pure pipe subgraph, contained in a general graph, can be shrunk to a single point. We cannot agree with this statement, since a subgraph can have many boundary points in which nodal P-variables are different, see Figure 4. The subgraph is not shrunk to a point, but to a generalized element containing N_b boundary points, or *pins* like in a microchip. We refer to such a generalized element further as *multipin*. As we show below, this element introduces not one, but N_b equations.

Without loss of generality, we can consider a connected subgraph for which the pins have definite flows serving as source/sink boundary conditions for the subgraph, as well as N_b nodal P-variables. One condition necessary for the stationary problem is the annulation of the sums of boundary flows. Here, for definiteness, we place all external sources/sinks in the subgraph, including Q_{set} and P_{set} nodes, on separate pins. Further, considering one of the boundary nodes as a point with a given pressure, the procedure from [11] uniquely reconstructs all other N_b-1 boundary pressures in terms of the first pressure and the boundary flows. The conditions for the equality of the reconstructed pressures to the given boundary pressures are the equations presenting the multipin for the external graph, totaling N_b equations.

In principle, it seems possible to precompute these N_b functions on a grid in the space of parameters and use fast interpolation algorithms to represent the multipin. The problem is the rapid increase of the grid data volume with the increasing dimension of N_b. In our approach, we have restricted our calculations to 2-pins, $N_b = 2$, which generally allows 2D tabulation (the pixel buffer from [8]). In this paper, we concentrate on the quadratic pipe model (1), which allows to encapsulate all the characteristics in one R-parameter and to perform all calculations analytically, without tabulated functions. Below, we consider also an intermediate case, where 1D-tabulation by splines is used. Thus, we avoid *curse-of-dimensionality* problems existing for general multipins and are still capable to reduce the dimension of the problem considerably.

In the remainder of this subsection, we consider in more detail an interesting question, why the multipin, regardless of its structure, is described by the same number of equations. Indeed, the number of equations external to the excluded subgraph is the same and does not depend on the topology of the subgraph. After eliminating the subgraph, the system must remain closed, meaning that the subgraph introduces the same number of equations. To calculate this number, it is enough to consider a specific configuration.

In Figure 5a, the star-like multipin is considered. One equation is the zero sum of the flows into the multipin. The Kirchhoff law in the center is equivalent to this equation. There is one P-variable in the middle, but there are also N_b conditions relating it to the boundary P_b and Q_b. In total, N_b-pin is equivalent to N_b equations on the boundary P and Q. Figure 5b shows the case when N_b is even and N_b-pin represents $N_b/2$ conditions for equality of incoming and outgoing flows, as well as $N_b/2$ of element equations. In total, we obtain N_b equations. In fact, even the connectivity of the graph is not important here. In Figure 5c, the case of an empty subgraph is considered, when all pins hang freely. Then, $Q_b = 0$ in all of them, comprising N_b equations.

b) Possible generalizations of friction laws: in the equations of the element, a general power dependence can be used, as was done in [11]. The consideration is quite similar. The element equation, series and parallel connections are described by:

$$P_{in}|P_{in}| - P_{out}|P_{out}| = RQ|Q|^{\alpha-1},\ \alpha \geq 1, \qquad (10)$$
$$R^s_{12} = R_1 + R_2,$$
$$R^p_{12} = \left(R_1^{-1/\alpha} + R_2^{-1/\alpha}\right)^{-\alpha}.$$

The quadratic law (1) corresponds to $\alpha = 2$.

Contraction of the leaf and reverse reconstruction are done in the same way.

Consider a more general case:

$$F(P_{in}) - F(P_{out}) = G(Q), \qquad (11)$$

where F, G are monotonously increasing functions, every element has an own G, while F is the same for all elements

(strictly speaking, it is enough if F is the same in a connected component of the graph).

For series connections, the equations can be combined as before:

$$F(P_1) - F(P_3) = G^s_{12}(Q), \qquad (12)$$
$$G^s_{12}(Q) = G_1(Q) + G_2(Q).$$

If the original functions were monotonic, then their sum will also be. The inverse reconstruction is:

$$P_2 = F_{inv}(F(P_1) - G_1(Q)), \qquad (13)$$

where by subscript *inv* we denote the inverse 1D-function, so as not to be confused with the algebraic inversion: $x^{-1} = 1/x$.

For parallel connections, the equations can be combined analogously:

$$F(P_1) - F(P_2) = G^p_{12}(Q), \qquad (14)$$
$$G^p_{12} = (G_{1,inv} + G_{2,inv})_{inv}.$$

Proof:

$$F(P_1) - F(P_2) = x = G_1(Q_1) = G_2(Q_2),$$
$$Q_1 = G_{1,inv}(x),\ Q_2 = G_{2,inv}(x),\ Q =$$
$$Q_1 + Q_2 = G_{1,inv}(x) + G_{2,inv}(x) = G^p_{12,inv}(x),$$
$$x = G^p_{12}(Q) = (G_{1,inv} + G_{2,inv})_{inv}(Q). \blacksquare$$

It can be seen that the resulting G-function is also monotonic. The structure of the formulas for quadratic and α-power resistance is also clear: the inverse of the power function is also a power function. Thus, the inverse reconstruction is:

$$Q_1 = G_{1,inv}(G^p_{12}(Q)),\ Q_2 = G_{2,inv}(G^p_{12}(Q)). \qquad (15)$$

To store 1D functions $y(x)$, one can use lists of tabulated values (x_n, y_n) and interpolate between them using cubic splines. Outside the working area $|P| \le 150$ bar, $|Q| \le 1000$ Nm^3/h, the data can be extended by linearly growing functions, similar to [6]. Such a representation is convenient for inverting the functions, for which it suffices to swap $(x_n, y_n) \to (y_n, x_n)$ and reconstruct the splines [16]. The accuracy of this procedure is controlled by the smoothness of the function and the density of subdivision. The computational complexity is proportional to the number of tabulated values, $O(N)$.

In the problems we are considering, the functions are odd: $y(-x) = -y(x)$. This means that it is enough for them to construct splines in the region $x \ge 0$ and use the symmetry for complete reconstruction. In addition, the functions have a vanishing derivative at zero, for example, $y = x|x| = x^2 \operatorname{sgn} x$, which leads to a non-smooth root dependence for inverse functions $x = \sqrt{|y|} \operatorname{sgn} y$. This leads to problems for representing such functions by cubic splines. In fact, as noted in [6], vanishing of the derivative also leads to instability of the solver. The case $Q = 0$ can occur in large regions of the network in the absence of a flow in them. This leads to zeroing of the derivative of the function $Q|Q|$ and entails the degeneration of the Jacobi matrix of the complete system. To overcome this problem, the laminar term $Q|Q| + \epsilon Q$ must be added to this function; similar regularizing terms must also be added to the P-functions. After this, the problem with the zero derivative disappears and does not hinder the spline inversion.

c) Precise friction laws: better precision can be achieved by Nikuradze and Hofer formulae [3][4]. These differential formulae can be analytically integrated under assumption of slow variation of temperature and compression factor over the pipe. If needed, the long pipes can be subdivided into smaller segments to achieve the necessary precision of the modeling. This piecewise integration approach is similar to the *finite element method* in modeling of flexible materials, flow dynamics, etc. The resulting formulae have the same quadratic form (1), with the resistance $R(Q, P_1, P_2)$ weakly (logarithmically) dependent on the flow and the pressures. Direct comparison between the quadratic and Hofer pipe laws on our test networks shows the difference on the level of 7-10%. The practical use of calculations with the approximate quadratic formula is a rapidly computable starting point for the subsequent refinement iterations with the precise formula. The gravitational term, available in the precise formula and taking into account the profile of the terrain, can also be embedded in the quadratic formula:

$$P_1|P_1|(1+\gamma) - P_2|P_2|(1-\gamma) = ... \qquad (16)$$
$$\gamma = \mu g(H_1 - H_2)/(R_{gas} T z),$$

where the dots denote the flow-dependent right part, in any form that we have considered. The dimensionless hydrostatic factor γ is determined by the gravitational acceleration g, the height difference $H_1 - H_2$ and the usual gas parameters. In real problems, the parameter γ is small, $|\gamma| \ll 1$, so the factors $(1 \pm \gamma)$ do not change the signature of the terms in the equation.

d) Inverse reconstruction: for practical purposes, it is enough to solve the problem on the reduced graph, the topological skeleton. The users are mainly interested in the values of flows and pressures at the end points of pipe subgraphs, where they are connected to active elements such as compressors and regulators or directed to the end consumers. One also needs to control the feasibility indicator $P > 0$. As we now show, it is enough to control this indicator at the endpoints.

Consider GSPG operations in the presence of nodes with negative pressure. For parallel connection, in the presence of negative pressure in the end node, it remains there after the reduction. For series connection, if there is negative pressure at the intermediate node, it will also be negative at the end node downstream. Indeed, considering the most general case with gravity corrections,

$$P_3|P_3|(1-\gamma) = P_2|P_2|(1+\gamma) - R_2 Q|Q|, \qquad (17)$$

since the factors $(1 \pm \gamma), R_2$ are positive, for $P_2 < 0$, $Q \ge 0$, we get $P_3 < 0$. Only contraction of a leaf with $Q_{set} > 0$ can be a problem, since this procedure can hide a negative pressure node downstream. As we have already explained, there is an option to block contracting leafs with nonzero Q_{set}. In this case, it suffices to check $P > 0$ at the end nodes of the pipe graph.

On the other hand, the data recovery in reduced elements is a straightforward analytical procedure. For this, a complete reduction history with all intermediate parameters and/or tabulated functions must be recorded. Then, the above-described inverse operations can be applied. On the graph obtained, it is possible to monitor the fulfillment of the condition $P > 0$ or the enhanced condition $P > 1$ bar or any other inequality on pressures and flows.

e) Level 3, Q_{set} movement algorithm: consider the two graphs depicted in Figure 6. Assuming that the central element is described by the general equation $F(P_1, P_2, Q)$, we require the equivalence of solutions, connecting these equations with the shift transformation of the argument:

$$F_b(P_1, P_2, Q) := F_a(P_1, P_2, Q + Q_{set}), \quad (18)$$
$$F_a(P_1, P_2, Q) = 0 \;\Rightarrow\; F_b(P_1, P_2, Q - Q_{set}) = 0.$$

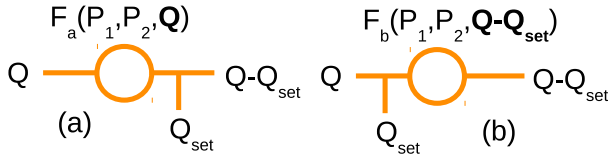


Figure 6. Q_{set} movement algorithm.

As a result, it is possible to move the Q_{set} specifier along a graph to an arbitrary place. For example, all Q_{set} specifiers can be moved to the P_{set} node, which should be present in each connected component of the graph. In this case, the undefined flow in this node will be shifted by the total Q_{set} in the subgraph. Alternatively, one can move all Q_{set} specifiers into one main consumer, who will represent all consumers in the subgraph. Note that such transformations change the distribution of flows in the graph, representing only a virtual distribution, which is visually unsimilar, but mathematically equivalent to the original one. To represent the result, of course, all the displaced Q_{set} specifiers must return to their places using inverse transformations. Note also that the argument shifts change the position of zero and violate the oddness of the functions. This requires to modify the tabulation algorithms; the easiest way is to consider all dependencies as monotonic functions of general form.

f) Not only pipes, combining 2D characteristic maps: after all pipe subgraphs are reduced to 2-pins, the functions can be transformed to a more general representation, in one of the equivalent forms:

$$Q = F(P_1, P_2), \;\; P_1 = F(P_2, Q), \;\; P_2 = F(P_1, Q). \quad (19)$$

All other elements, such as compressors and regulators, can be represented in the same way. Such a representation can use the 2D-tabulation (pixmaps) algorithms described in [8], as well as piecewise linear monotone extensions outside of the working region. As a result, GSPG reduction can be continued at the level of 2D functions. Thus, our proposed strategy is to keep the low-dimensional representations as long as possible, such as quadratic equations or 1D-splines for pipes, and then, after the network is strongly reduced, proceed to pixmaps.

V. Improved algorithms

g) Triangulation: yet another possibility [8] is the representation of element equation as triangulated surface in the space of main variables (P_1, P_2, Q), see Figure 7. Such a surface defines a continuous piecewise linear function, e.g., $P_2(P_1, Q)$, whose monotony conditions are related with the direction of normal at each triangle. Namely, the normal n should be directed in octant, corresponding to $(+, -, -)$ signature. This property can be exhaustively checked for all triangles. Practically, the marginal signatures, e.g., $(+, -, 0)$,

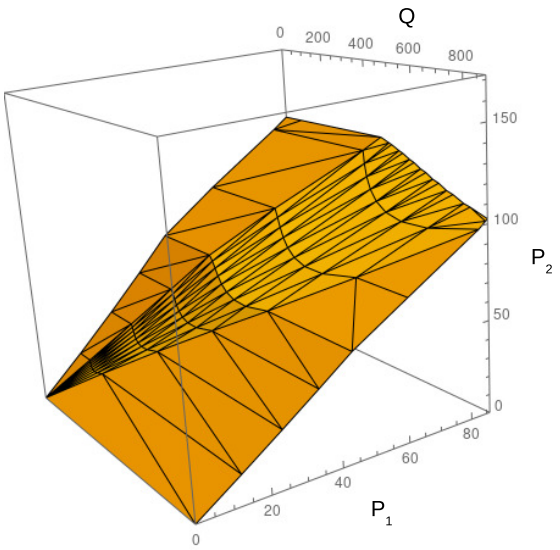


Figure 7. An example of triangulated representation $P_2(P_1, Q)$.

as for some triangles in the figure, are also allowed, if one adds a linear function $\epsilon(P_1 - P_2 - Q)$, with small positive regularization parameter ϵ, to the element equation.

When specified in a box on (P_1, Q) plane, the element equation can be continued outside of the box by the formula [6]:

$$\begin{aligned}
&\tilde{P}_2(P_1, Q) = P_2(\hat{P}_1, \hat{Q}) \quad (20)\\
&+k_P(\min(P_1 - P_{min}, 0) + \max(P_1 - P_{max}, 0))\\
&+k_Q(\min(Q - Q_{min}, 0) + \max(Q - Q_{max}, 0)),\\
&\hat{P}_1 = \min(\max(P_1, P_{min}), P_{max}),\\
&\hat{Q} = \min(\max(Q, Q_{min}), Q_{max}),
\end{aligned}$$

with constants $k_P > 0$, $k_Q < 0$. It is a particular case of more general formula:

$$\begin{aligned}
&\tilde{F}(x_1, ..., x_n) = F(\hat{x}_1, ..., \hat{x}_n) \quad (21)\\
&\quad + \sum_{i=1}^{n} k_i(\min(x_i - a_i, 0) + \max(x_i - b_i, 0)),\\
&\hat{x}_i = \min(\max(x_i, a_i), b_i)
\end{aligned}$$

for continuation of the function defined in a box in R^n, monotone with respect to each argument, to the outside, preserving the signature $\mathrm{sgn}(k_1, ..., k_n) = \mathrm{sgn}(\nabla F)$.

In this way, we obtain a piecewise linear representation of the function, fulfilling all necessary monotony conditions in the whole R^2 domain. As a result, the system satisfies the conditions [17] for the existence of solution of a system $f(x) = c$ for piecewise-linear function f and arbitrary c. This solution can be found using Katzenelson algorithm [18][17][5], converging to the solution in a finite number of steps. As an alternative, one can use a standard Armijo rule [19], applicable for smooth functions, which for the considered piecewise linear functions also converges to the solution in a finite number of steps. The reason for this is that the both algorithms require a finite number of steps to find a triangle where the solution is located, then, due to the linearity of the functions, converge to the solution in one iteration.

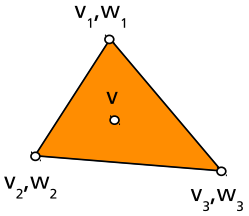


Figure 8. Illustration to barycentric coordinates.

Considering this representation in the context of topological reduction, we see that the necessary lookup operation $P_2^{(2)}(P_2^{(1)}(P_1,Q),Q)$ for serial connection is straightforward. For parallel connection, consider the formula (20), representing the element equation as continuous piecewise linear function on the whole plane R^2. It can also be reprojected to the other axes, e.g., $Q(P_1,P_2)$. Then we make a new triangulation with the required resolution on a square (P_1,P_2), and evaluate Q in the vertices of the triangles. Repeating the same operations with the parallel element, sum Q_1+Q_2 in the vertices of the shared triangulation and continue outside of the square by the $Q(P_1,P_2)$-analog of the formula (20). It can be then reprojected back to $P_2(P_1,Q)$-representation if needed. The operation count for the reduction, valid for both serial and parallel connections, is $O(N_{tri})$ multiplied to the cost of one evaluation of the triangulated function representation. This cost varies between $O(N_{tri})$ if direct search of triangle is implemented in the function evaluation and $O(1)$ for indexed search.

h) Low level implementation: construction of triangulated surface starts with a subdivision, done either with experimentally measured points or with continuous models, sampled with the necessary precision. This is done once in a preprocessing stage and the list of triangles is stored. Further, during the solution stage, for a given point $(v^x,v^y)=(x,y)=(P_1,Q)$ the triangle is found, containing this point. The search is done either directly, or with creating an auxiliary indexing structure during the preprocessing. This structure can be, e.g., a rectangular grid, assigning a sublist of triangles to a particular cell. After the triangle is found, its representation in barycentric coordinates is used, see Figure 8. Namely, the weights w_i are introduced at the vertices v_i of the triangle, $i=1..3$, such that

$$\sum_i w_i=1,\ v^k=\sum_i w_i v_i^k,\ k=x,y. \tag{22}$$

The same weights are used for linear interpolation of the function $v^z=z=P_2$ we are looking for:

$$z=\sum_i w_i z_i. \tag{23}$$

Solving (22) for w_i and substituting it to (23), obtain:

$$\begin{aligned}
&w_i=l_i^0+l_i^x x+l_i^y y,\ i=1..3, \qquad (24)\\
&z=l_4^0+l_4^x x+l_4^y y,\\
&l_i^k=n_i^k/d,\ k=0,x,y,\ i=1..4,\\
&n=\{\{x_3y_2-x_2y_3,-y_2+y_3,x_2-x_3\},\\
&\{-x_3y_1+x_1y_3,y_1-y_3,-x_1+x_3\},\\
&\{x_2y_1-x_1y_2,-y_1+y_2,x_1-x_2\},\\
&\{x_3y_2z_1-x_2y_3z_1-x_3y_1z_2\\
&+x_1y_3z_2+x_2y_1z_3-x_1y_2z_3,\\
&y_3(z_1-z_2)+y_1(z_2-z_3)+y_2(-z_1+z_3),\\
&x_3(-z_1+z_2)+x_2(z_1-z_3)+x_1(-z_2+z_3)\}\},\\
&d=x_3(-y_1+y_2)+x_2(y_1-y_3)+x_1(-y_2+y_3).
\end{aligned}$$

The necessary and sufficient condition for the point to belong to the triangle is that all $w_i\geq 0$. Note also that the function z in the triangle is linear with respect to (x,y), and its gradient is given by

$$(\partial z/\partial x,\partial z/\partial y)=(l_4^x,l_4^y). \tag{25}$$

i) Existence and uniqueness of solution: mathematically, the difference between the continuous piecewise linear and smooth modeling in our class of problems is that the mapping used in the equation $f(x)=c$ is either invertible continuous (homeomorphism) or invertible differentiable, for both forward and backward mappings (diffeomorphism). Any smooth mapping can be approximated by the piecewise linear one. For the considered class of element equations, any piecewise linear representation can be approximated by a smooth surface with the same signature of the normal. For this purpose, one should simply smooth the corners for the polyhedrons in 3 dimensions, representing our element equations. In paper [6] we have proven that if the element equations in the form $F(P_1,P_2,Q)=0$ everywhere possess a signature of the gradient $\nabla F=(+,-,-)$ and all connected components of the network have P_{set}-entries, then the Jacobi matrix $J=\partial f/\partial x$ of the system $f(x)=c$, composed of element and Kirchhoff equations, is globally non-degenerate. According to [19], the boundness $||J^{-1}||<C$ provides convergence of Newton method with Armijo line search rule, being started from an arbitrary point, therefore, the solution of the system $f(x)=c$ always exists. Also, non-degeneracy of the Jacobian means that the mapping $f(x)$ does not have folds or other singularities, as a result, the system $f(x)=c$ at all c possesses the same number of solutions, preimages of c under mapping f. This does not mean yet that there is a unique solution/preimage. Examples of non-trivial topology can be constructed, possessing everywhere non-degenerate Jacobian and having two or more preimages, e.g., mapping of torus to itself with $n>1$ winding numbers. Such examples can also be constructed for continuous piecewise linear mappings. To obtain a unique solution in the considered R^n case, we need to specify the behavior of the mapping at infinity. Below, we will show that in the considered problem, the element equation can be continued to infinity as a linear function with the necessary signature. As a result, the mapping $f(x)$ at infinity will be linear with non-degenerate Jacobian. Such a mapping with necessity has a single preimage, then due to the absence of singularities, there will be a single preimage of the mapping $f(x)$ everywhere. Thus, we prove that the mapping $f(x)$ is true diffeomorphism and the system $f(x)=c$ for the considered class of problems possesses *a unique solution*.

Lemma: the element equation $F(P_1,P_2,Q)=0$ with smooth F of signature $\nabla F=(+,-,-)$ in a box can be continued to infinity as a linear function, keeping the same signature everywhere.

Proof: at first, consider the continuation (21) of the function F, a linear function $F_0=\sum_i k_i x_i$ and linear interpolation between them: $F_1=(1-w)F+wF_0$, where $w(x)\in[0,1]$

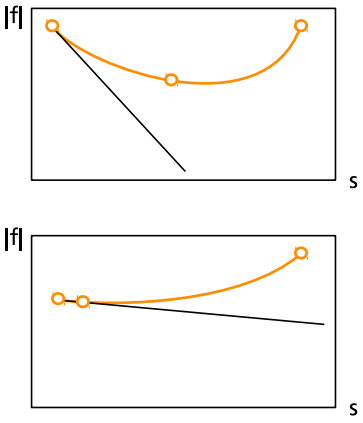


Figure 9. Illustration to Armijo rule. At the top: normal situation, at the bottom: stagnation of the algorithm.

is the weight function. Consider two spheres $|x| = R_{1,2}$, the first one contains the box of original F definition, the second one is of a greater radius $R_2 > R_1$. The function $w(x)$ is varied smoothly from 0 inside R_1-sphere to 1 outside R_2-sphere. Computing the gradient:

$$\nabla F_1 = (1-w)\nabla\tilde{F} + w\nabla F_0 + \nabla w(F_0 - \tilde{F}). \qquad (26)$$

The sum of the first two terms has correct signature automatically. For the third term, $|F_0 - \tilde{F}| < C$ is bounded, due to the equal slopes k_i in the linear function F_0 and the continuation $\tilde{F}$. By choosing $R_2 \gg R_1$, the function $w(x)$ can be selected, so that $|\nabla w|$ will be arbitrarily small. As a result, the third term can be made arbitrarily small, so that it will not change the signature of the gradient, defined by the sum of the first two terms. ■

Note that if the function F everywhere satisfies monotony conditions, it becomes unimportant, where $[a_i, b_i]$ limits of the box are set and how R_1 and R_2 spheres are selected. For every choice, the uniqueness of solution becomes proven. Therefore, the continuation beyond the box and interpolation between R_1 and R_2 spheres can be considered as auxiliary elements of the construction, provided for the purpose of proof only.

In more general terms, in differential topology, a degree of mapping $f(x) = c$ is defined as a sum $\deg f = \sum \operatorname{sgn} \det J$, taken over all preimages of a point c. This characteristics does not depend on the choice of a point c (assuming that the point is regular, $\det J \neq 0$). Actually, we have proven that at infinity $\deg f$ is equal $+1$ or -1, dependently on the sign of Jacobian at the linear continuation of the mapping. Then, it is the same everywhere, and since the Jacobian always has the same sign, the number of preimages is always 1, the system has a unique solution.

j) A choice of Newtonian stabilizer: as already mentioned, a combination of Newton method with Armijo line search rule [19] has guaranteed convergence in our problem class. A similar proof is also available for Katzenelson algorithm [17]. Armijo rule serves as a stabilizer to Newton method, which does not allow to perform too large steps, accepting only the steps which reduce the norm of the system residual. Katzenelson algorithm does the same for piecewise linear systems, testing whether the solution is located in a current piece, and if not, going along the Newton direction to the adjacent piece. Currently, we consider the mixed system of linear, piecewise linear and non-linear equations and prefer more generic Armijo rule. In certain cases, this leads to the following problem. Some of our elements (compressors, regulators in gas transport networks) possess degenerate equations, with zero signature instead of the prescribed ± 1. Such equations are regularized, generally, adding a small linear function with a definite signature. However, numerically these systems are almost degenerate. In particular, we observe a stagnation of iterations when the problematic region is approached. The reason is that Armijo rule attempts to reduce the function, which is almost constant in such regions, performing *a plateau optimization*, see Figure 9. Consider normalized Newton direction $n = dx/|dx|$, $dx = -J^{-1}f$. A variation of the norm of residual, in linear approximation, in that direction is $d|f| = f^T Jns/|f|$, where s is the distance along n. As a result, $d|f| = -|f|s/|dx|$. Approaching the degenerate region, in general position, $|dx| \to \infty$, thus, $d|f| \to -0$, in linear approximation the norm of residual along Newton direction is almost constant. This leads to stagnation of iterations. Indeed, $\Delta|f| = c_1 s + c_2 s^2 + ...$, a minimum of this function of s is located at $s^* = -c_1/(2c_2)$, at $c_1 < 0$ and $c_2 > 0$. Since $c_1 \to -0$, any non-zero quadratic term will lead to the vanishing step $s^* \to 0$. Practically, being stopped at the given number of iterations, the stagnation of the algorithm is equivalent to its divergence.

To overcome this issue, we propose a relaxed version of Armijo algorithm, which recognizes degenerate situation and allows for larger steps in the problematic region.

Algorithm (relaxed Armijo rule):

```
repeat until convergence:
  do Newtonian step dx
  set λ = 1
  trial point: x_t = x + λdx
  do λ = λ/2                                  (* bisection *)
    if |f(x)|/|dx| < ξ_0
    then cond = (|f(x_t)| < |f(x)| + df)
    else cond = (|f(x_t)| < (1 − αλ)|f(x)|)
  until cond                                  (* exit condition *)
  x = x_t                                     (* trial point accepted *)
```

The algorithm detects plateau situation by estimating of the lowest SVD eigenvalue (Singular Value Decomposition):

$$Jdx = -f, \; n = dx/|dx|, \qquad (27)$$
$$\xi^2_{min} \leq \xi^2 = n^T(J^T J)n = |f|^2/|dx|^2.$$

It has two parameters: ξ_0 is the upper threshold for the estimated lowest eigenvalue, df – the allowed increase of the residual in plateau situation. The exit condition is therefore modified: for $\xi < \xi_0$ a slight increase of the norm of the residual is allowed, not more then by df. Otherwise, a standard sufficient decrease rule is applied.

Our tests show that this modification of the algorithm drastically improves the convergence rate, for 60 tested networks the old algorithm brings 12 scenarios to divergence, while the

new relaxed version diverges only in 1 case, where it causes the Newton iteration to cycle. Switching the relaxation off for this particular scenario helps to achieve 100% convergence.

In the considered almost degenerate cases none of the algorithms, standard or relaxed Armijo rule, provide the guaranteed convergence. Empirically, the relaxed version behaves better. A construction of stable algorithm for processing almost degenerate cases remains a challenge for further developments.

k) Coupling to other sectors: note that topological reduction is possible not only for gas transport networks, but also for other network problems with element equations with the power law Q^α or other from the considered forms above. A practically important problem is a coupling of transport networks from different energetic sectors, e.g., gas, water, electricity, etc. In our previous work [7] it was shown how the coupling of sectors can be done on the modeling level. The question of stability and global convergence of solution for multi-sectoral case has been insufficiently studied. However, we can already identify two cases where the solution procedure is globally convergent.

Case A: consider two sectors, both satisfying the conditions of global convergence from [6]. If two sectors are not coupled, their Jacobi matrix possesses block-diagonal structure. The coupling produces terms located in non-diagonal blocks. Jacobian of the whole system depends continuously on these terms and is non-degenerate when these terms are switched off. Therefore, when the coupling between the sectors is sufficiently weak, then the Jacobian remains non-degenerate. In practice, the usage of this property can encounter one obstacle. If the element equation has marginal signature of derivatives, the Jacobian is almost degenerate and the weak coupling can lead it to the complete degeneracy.

Case B: consider, for definiteness, the coupling of gas transport and water thermal sectors, see Figure 10. The modeling of water thermal sector has been described in [7]. In the simplest case, it is represented by Kirchhoff law for thermal energy and element equations of the form:

$$\sum_e c_v Q_e T_e = 0,\ T_{n1} + \epsilon T_{n2} = T_e,\ Q_e > 0. \qquad (28)$$

The temperature in the edge T_e is defined by the temperature T_{n1} in the node upstream, also a small reverse diffusive term is introduced for regularization. In the presence of the given T_{set} temperature per every connected component of the graph, the problem is linear and non-degenerate. Moreover, the problem belongs to the already described class [7] with (linear) Kirchhoff nodal equations and (linear) element equation with (marginally) correct signature. For the coupling of sectors, an element is introduced representing a combustion chamber and a heat exchanger. Their equation

$$\Delta E = \eta H_m Q_g = c_v Q_w (T_{n2} - T_{n1}), \qquad (29)$$

in fact, transforms the energy of combustion of gas $H_m Q_g$, with a certain efficiency coefficient η, to the thermal energy of water $c_v Q_w \Delta T$. In this case, the definition of flows in one sector, gas or water thermal, can be renormalized, for the purpose of proof only, so that the energy from one sector will flow directly to another sector. In this way, a combined problem will be formulated, with Kirchhoff law for the energy flow and all elements possessing correct signature.

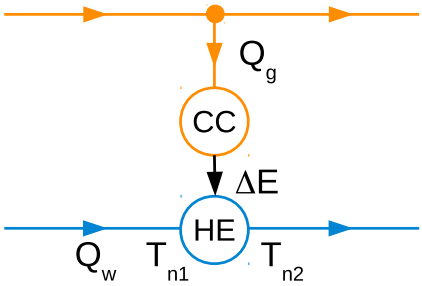


Figure 10. An example of coupling between gas and water thermal sectors. Gas (orange), water thermal (blue), combustion chamber (CC), heat exchanger (HE).

Therefore, for the combined problem the global convergence can be proven, without even using the fact that the water part of the system is linear. The generalization to the case of multiple contact points between sectors with different transfer coefficients is still to be studied. An interesting property of the considered formulation of multi-sectoral problems in the field of energetics is that all flows participating in Kirchhoff equations are energetic and universal, while the other type of flows, e.g., mass flow, molar flow, volume flow, etc., are related to the energetic ones by proportionality coefficients, depending on the particular sector.

VI. Conclusion

In this paper, the general method of topological reduction for the network problems has been presented, using gas transport networks as an example. The method uses a contraction of series, parallel and tree-like subgraphs, containing the edge elements, described by quadratic, power law or general monotone dependence. This way, we achieve the goal of significant lossless reduction of the graphs and we accelerate solution procedure correspondingly. A large set of realistic network examples of different complexity have been used for the benchmarking of the method. The statistical distribution for the reduction and the acceleration factors has been evaluated, together with corresponding minimum, maximum, median and mean values. Comparing with the original network (level0), the elimination of superconductive elements and disconnected parts (level1) brings the reduction factor into the range 1.9-2.9, further GSPG reduction with fixed Qsets (level2) multiplies it by the factor 1.26-2.9, then GSPG reduction with moving Qsets (level3) gives a projected multiplicative factor 1.6-2.3. We have done performance comparison between the numerically implemented levels 1, 2. While level1 is absolutely necessary for the convergence, level2 brings the acceleration factor varied from 0.34 (deceleration) to 61.4 (strong acceleration), with median 1.54 and mean 2.71, for the solution procedure, with a little overhead for GSPG pre-filtering.

The possible extensions of the method include iterative schemes for Nikuradze and Hofer formulae, rapid inverse reconstruction of data in reduced subgraphs, Qset movement algorithm for deeper reduction and the extension of the reduction methods to other elements using 2D tabulation (pixmaps). Additional extensions have been proposed, including triangulated element equations, continuation of the equations at infinity, providing uniqueness of solution, a choice of Newtonian stabilizer for nearly degenerated systems. The generalization

of the method for various sectors in the field of energetics has been proposed, including gas networks, water networks, electric networks, as well as for coupling of different sectors.

VII. Acknowledgment

We are grateful to the organizers and participants of INFOCOMP 2019 conference for fruitful discussions. The work has been supported by the German Federal Ministry for Economic Affairs and Energy, project BMWI-0324019A, MathEnergy: Mathematical Key Technologies for Evolving Energy Grids and by the German Bundesland North Rhine-Westphalia using fundings from the European Regional Development Fund, grant Nr. EFRE-0800063, project ES-FLEX-INFRA.

References

[1] A. Baldin et al., "Topological Reduction of Gas Transport Networks", in Proc. of INFOCOMP 2019, IARIA, 2019, pp. 15-20.

[2] J. Mischner, H.G. Fasold, and K. Kadner, System-planning basics of gas supply, Oldenbourg Industrieverlag GmbH, 2011 (in German).

[3] M. Schmidt, M. C. Steinbach, and B. M. Willert, "High detail stationary optimization models for gas networks", Optimization and Engineering, vol. 16, 2015, pp. 131-164.

[4] T. Clees, "Parameter studies for energy networks with examples from gas transport", Springer Proceedings in Mathematics & Statistics, vol. 153, 2016, pp. 29-54.

[5] T. Clees, N. Hornung, I. Nikitin, and L. Nikitina, "A Globally Convergent Method for Generalized Resistive Systems and its Application to Stationary Problems in Gas Transport Networks", In Proc. SIMULTECH 2016, SCITEPRESS, 2016, pp. 64-70.

[6] T. Clees, I. Nikitin, and L. Nikitina, "Making Network Solvers Globally Convergent", Advances in Intelligent Systems and Computing, vol. 676, 2017, pp. 140-153.

[7] A. Baldin et al., "Universal Translation Algorithm for Formulation of Transport Network Problems", in Proc. SIMULTECH 2018, vol. 1, pp. 315-322.

[8] T. Clees, I. Nikitin, L. Nikitina, and Ł. Segiet, "Modeling of Gas Compressors and Hierarchical Reduction for Globally Convergent Stationary Network Solvers", Int. J. On Advances in Systems and Measurements, IARIA, vol. 11, 2018, pp. 61-71.

[9] D. Eppstein, "Parallel recognition of series-parallel graphs", Information and Computation, vol. 98, 1992, pp. 41-55.

[10] N. M. Korneyenko, "Combinatorial algorithms on a class of graphs", Discrete Applied Mathematics, vol. 54, 1994, pp. 215-217.

[11] R. Z. Rios-Mercado, S. Wu, L. R. Scott, and E. A. Boyd, "A Reduction Technique for Natural Gas Transmission Network Optimization Problems", Annals of Operations Research, vol. 117, 2002, pp. 217-234.

[12] T. Clees et al., "MYNTS: Multi-phYsics NeTwork Simulator", In Proc. SIMULTECH 2016, SCITEPRESS, pp. 179-186.

[13] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming", Mathematical Programming, vol. 106, 2006, pp. 25-57.

[14] Mathematica, Reference Manual, http://reference.wolfram.com [retrieved: June, 2019].

[15] MATLAB, https://de.mathworks.com/products/matlab.html [retrieved: June, 2019].

[16] T. Clees, I. Nikitin, L. Nikitina, and S. Pott, "Quasi-Monte Carlo and RBF Metamodeling for Quantile Estimation in River Bed Morphodynamics", Advances in Intelligent Systems and Computing, vol. 319, 2014, pp. 211-222.

[17] M. J. Chien and E. S. Kuh, "Solving piecewise-linear equations for resistive networks", International Journal of Circuit Theory and Applications, vol. 4, 1976, pp. 1-24.

[18] J. Katzenelson, "An algorithm for solving nonlinear resistor networks", Bell System Technical Journal, vol. 44, 1965, pp. 1605-1620.

[19] C. T. Kelley, Iterative Methods for Linear and Nonlinear Equations, SIAM, Philadelphia, 1995.

Parameter Identification and Model Reduction in the Design of Alkaline Methanol Fuel Cells

Tanja Clees[1,2], Bernhard Klaassen[2], Igor Nikitin[2], Lialia Nikitina[2], Sabine Pott[2], Ulrike Krewer[3], Theresa Haisch[3,4], Fabian Kubannek[3]

[1] University of Applied Sciences Bonn-Rhein-Sieg, Sankt Augustin, Germany
[2] Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany
Email: Name.Surname@scai.fraunhofer.de
[3] Institute of Energy and Process Systems Engineering, Technical University, Braunschweig, Germany
Email: {u.krewer, f.kubannek}@tu-braunschweig.de
[4] DECHEMA Research Institute, Frankfurt am Main, Germany
Email: theresa.haisch@dechema.de

***Abstract*—Alkaline methanol oxidation is an important electrochemical process in the design of efficient fuel cells. Typically, a system of ordinary differential equations is used to model the kinetics of this process. The fitting of the parameters of the underlying mathematical model is performed on the basis of different types of experiments, characterizing the fuel cell. In this paper, we describe generic methods for creation of a mathematical model of electrochemical kinetics from a given reaction network, as well as for identification of parameters of this model. We also describe methods for model reduction, based on a combination of steady-state and dynamical descriptions of the process. The methods are tested on a range of experiments, including different concentrations of the reagents and different voltage range.**

Keywords–modeling of complex systems; observational data and simulations; advanced applications; mathematical chemistry.

I. INTRODUCTION

This work extends our conference paper [1], where the methods for model reduction in electrochemical kinetics of alkaline methanol fuel cells have been introduced. Here we present more details on different types of experiments, characterizing the fuel cell behavior, as well as the methods for automatic generation of a mathematical model from chemical reaction description, the methods of parameter identification, and more testing results for these methods.

Galvanic cell is a chemical source of electric current based on interaction of two metals in electrolyte. Fuel cells are similar to galvanic cell but the reagents can be refilled multiple times. A special class of the fuel cells are direct fuel cells, in which formation of hydrogen is avoided, providing more safety for the usage. Renewable sources of energy are provided by alcohols, most commonly used are methanol and ethanol. In fuel cells, the chemical reaction of oxidation, the burning of the alcohols, is directly transferred to the electric energy.

The methods of parameter identification in electrochemical kinetics have been presented in our papers [2–4]. They include three types of experiments, measuring stationary state of the cell for a given voltage, small harmonic oscillations near this state and the response of the cell to a large amplitude variation of the voltage. The theoretical basis for these experiments can be found in the works [5–9]. Our purpose is the application of these general methods to a particular system, describing electrooxidation of the methanol in alkaline medium.

In Section II we describe the experiments, characterizing the fuel cell of alkaline methanol type. In Section III the methods of model generation and parameter identification are presented. Further improvements of the methods by model reduction and the results of their application to the experimental data are given in Section IV.

II. THE EXPERIMENT

The experimental setup is shown on Figure 1. It is efficiently designed to avoid external influence and minimize processes, which could violate the results of modeling. The cell itself (1) is made of teflon to provide maximal tightness of the parts. The working electrode (2) is a disk coated with a platinum catalyst. The disk is rapidly rotated to suppress diffusion effects. There is a counter electrode (3) made of a platinum wire and a reference electrode (4) made of Hg/HgO, used for recalculation of potentials. The temperature in the cell is measured by sensor (5). The forming and remaining environmental CO_2 is removed from the cell by permanent argon blow (6). The setup is put under deep vacuum. During the experiments, the voltage in the cell is set and the current is measured. Inbetween the experiments, high amplitude voltage variations are used for electrochemical cleaning of the electrode.

Three types of experiments are performed, see Figure 2. Polarization Curve (PC) represents the stationary state of the cell. For this purpose the voltage is changed very slowly in equidistant steps, on each step the stabilized current is measured. Electro-Impedance Spectrometry (EIS) probes the cell with harmonic oscillations of small amplitude in the vicinity of the stationary state, with the frequency varied in a large range. A linearized behavior of the system is characterized by complex-valued resistance, an impedance of the cell, which is computed and displayed on Nyquist plot diagrams. Cyclic Voltammetry (CV) probes the cell with a periodic saw-like voltage profile of high amplitude and measures the corresponding volt-ampere dynamical characteristics. The corresponding plot contains non-split stationary PC-part and a hysteresis loop, representing non-stationary, dynamical effects.

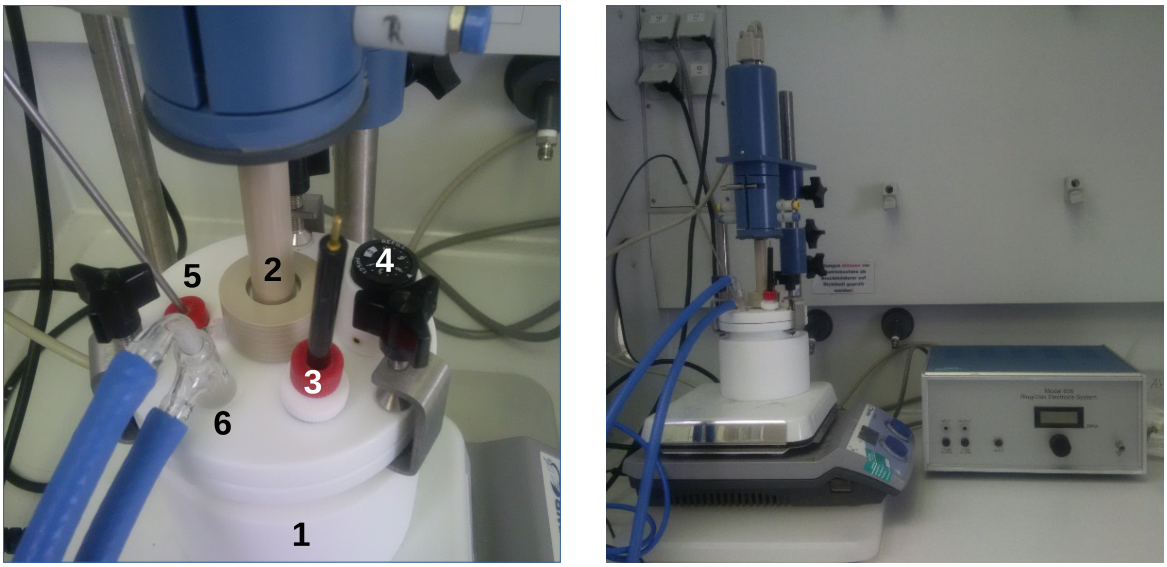


Figure 1. On the left: the experimental setup, consisting of a teflon cell (1) under deep vacuum, the rotating working electrode (2), the counter electrode (3), the reference electrode (4), the temperature sensor (5) and argon blow supply (6). On the right: a general view with service equipment.

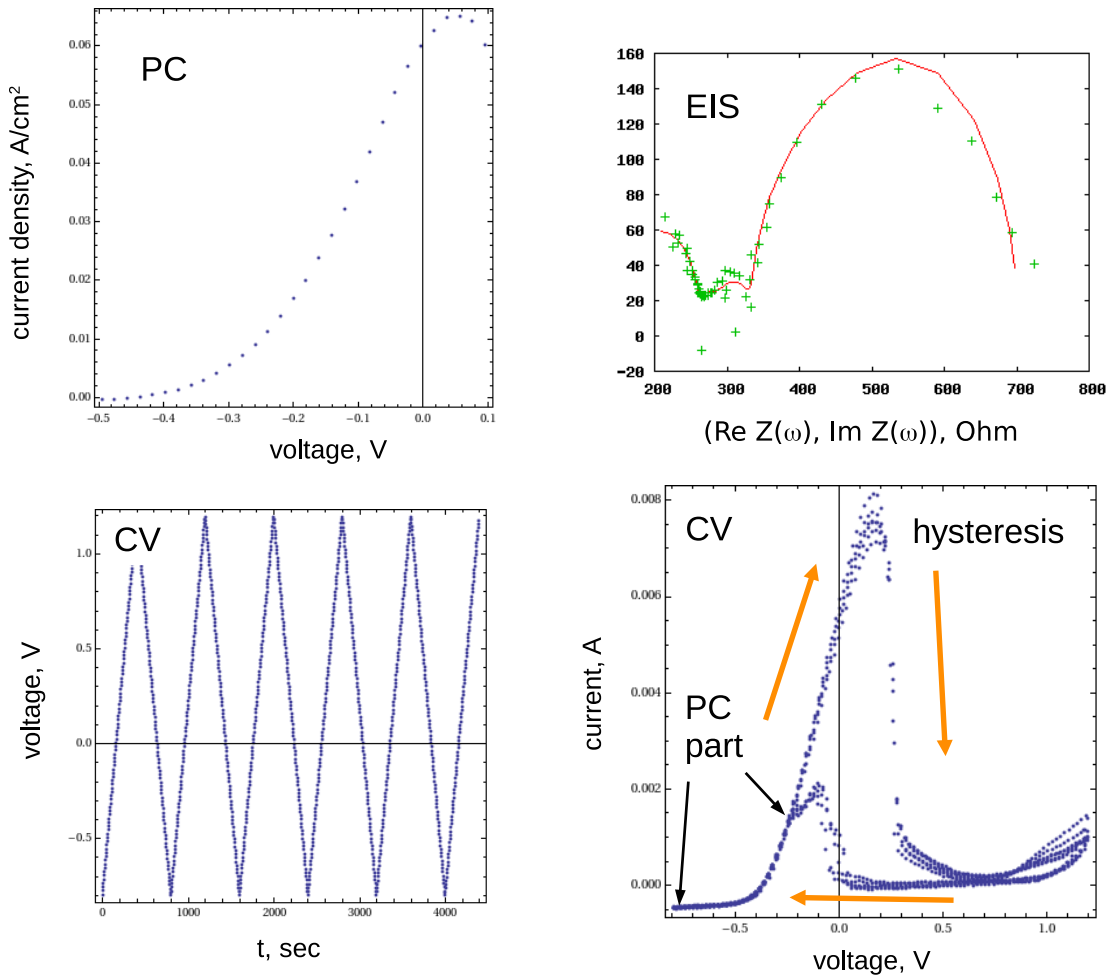


Figure 2. Three types of experiments (PC, EIS, CV), characterizing the cell.

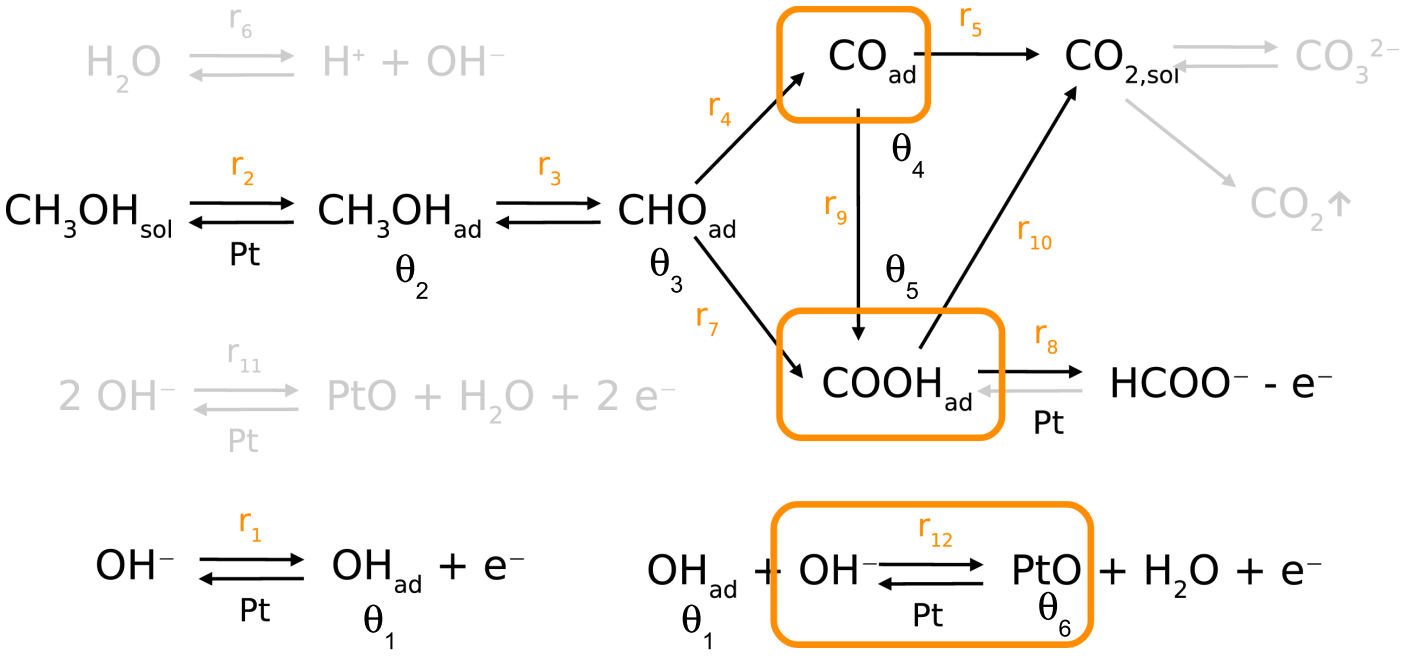


Figure 3. The chemical reactions network, the orange boxes show the reactions potentially responsible for the hysteresis effect on the CV plot.

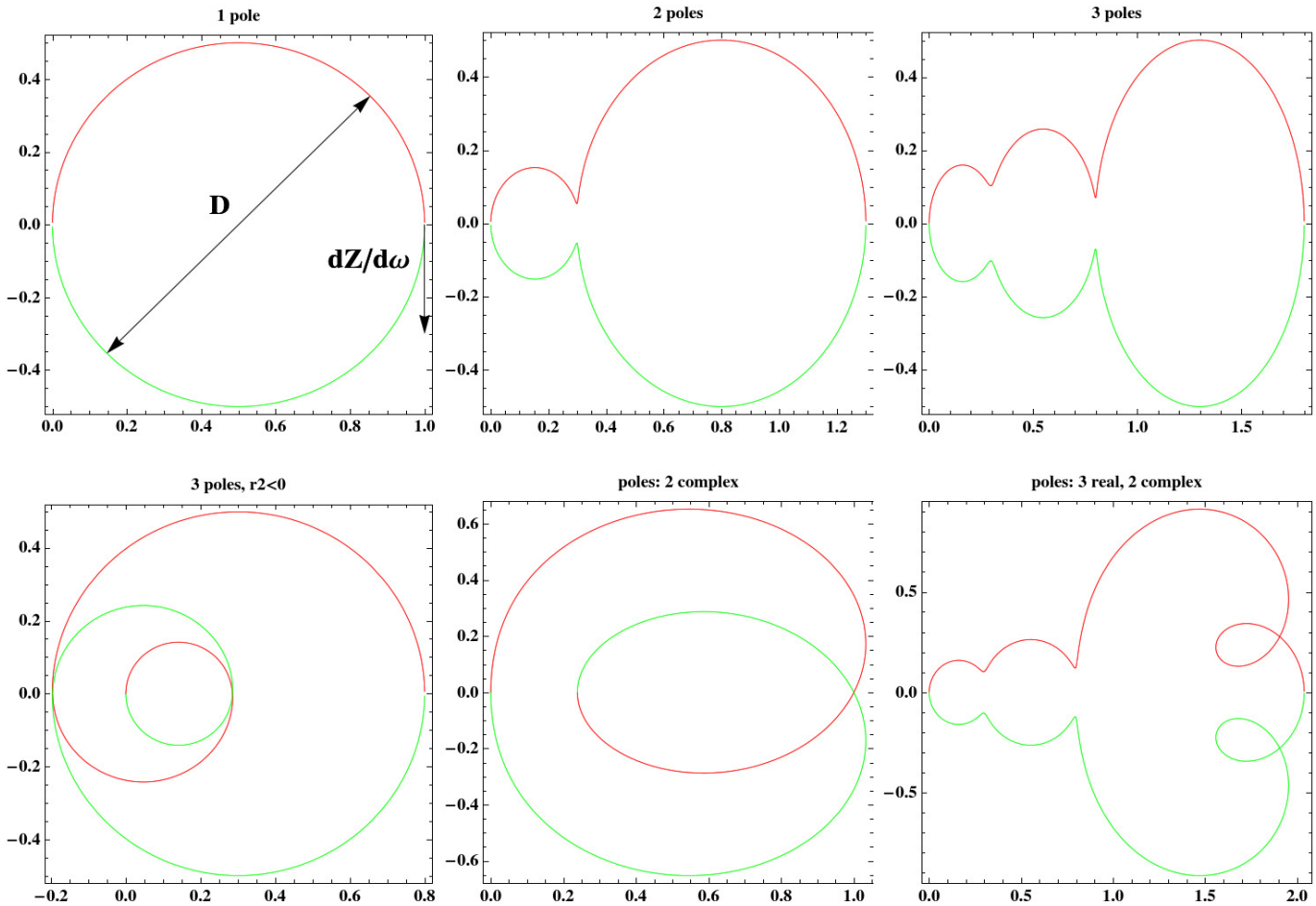


Figure 4. Typical shapes of EIS diagrams. Image from [4].

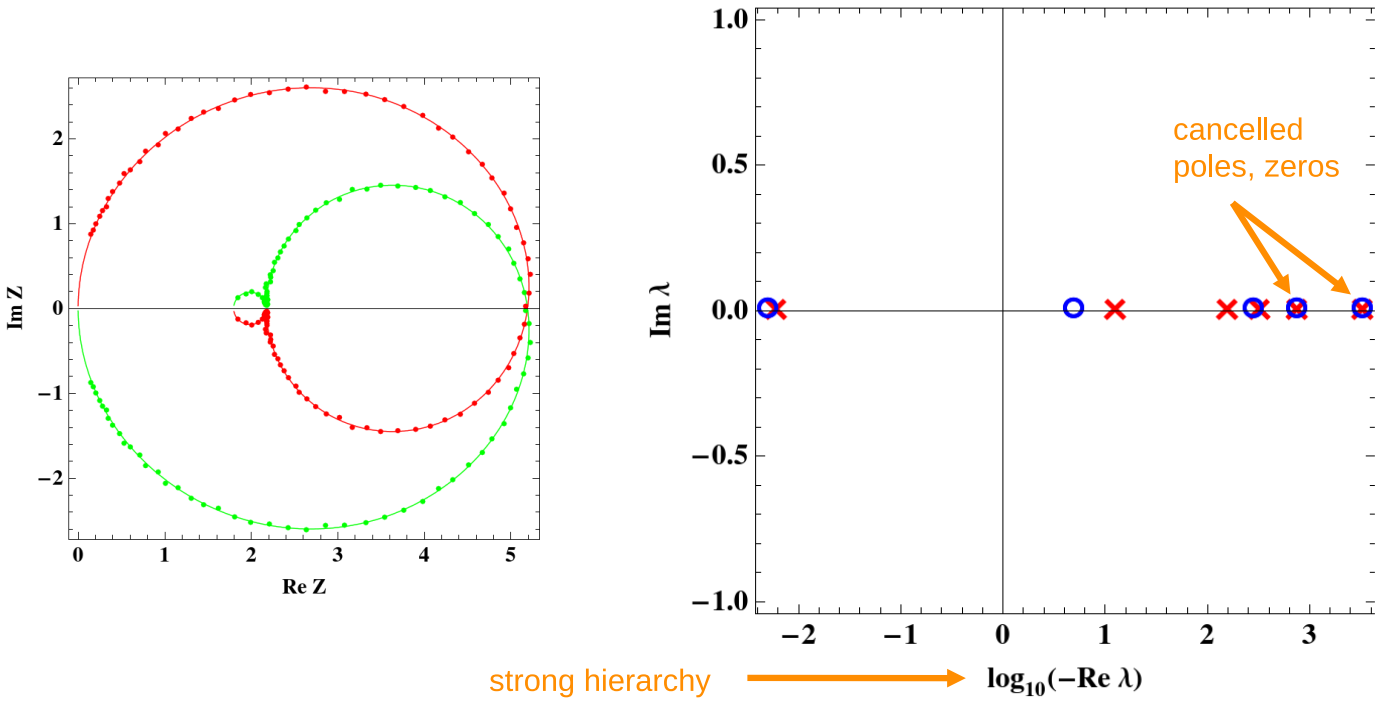


Figure 5. Nyquist plot (on the left) and reconstructed poles and zeros (on the right). Image from [4].

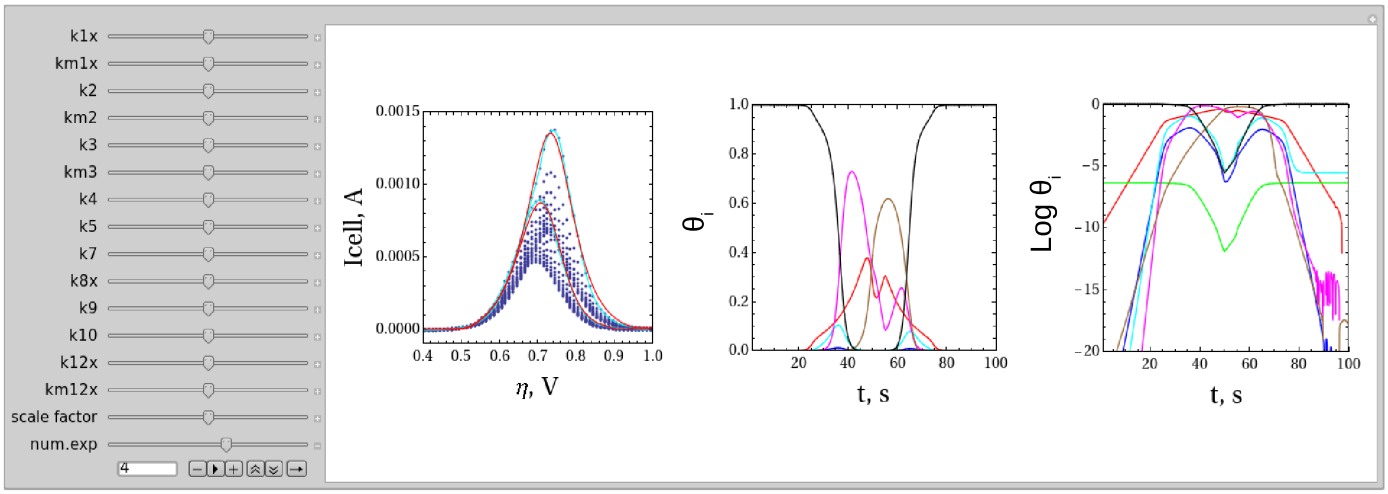


Figure 6. Interactive tool for parameter space exploration (in Mathematica v11).

III. The Model

The chemistry of the cell is described by the network of reactions, shown on Figure 3 and Table I. It includes the chain of the oxidation of carbon containing reagents, starting from methanol CH_3OH in the solution (subscript sol) and up to carbon dioxide CO_2, finally removed from the system. In parallel, the process of transformation of hydroxyl ion OH^- is present. All processes start from adsorption of fuel and hydroxyl on the platinum catalyst (subscript ad). Some of the reactions are reversible, so that the direct and reverse chemical processes are shown by two oppositely directed arrows. Our previous investigation [3] allows to exclude some originally postulated reactions from the model, due to their vanishingly small influence on the result. On Figure 3, these reactions are grayed out and in the mathematical model they can be omitted by setting the corresponding reaction constants to zero. Also in our study [2], [3] we have identified reactions, related with weakly coupled intermediates, which could serve as a source for the hysteresis and other dynamical effects. On the reaction scheme, they are highlighted by orange bubbles. Most of important among them is a reaction r_{12}, describing formation of platinum monoxide PtO.

Proceeding to the mathematical description of electrochemical kinetics, we note that the reaction network on Figure 3

can be represented as *a hypergraph*, a generalization of graph where an edge can join any number of vertices. It can be described by an incidence matrix a_{ij}, for ending vertices (j) entering in an edge (i). We use an approach, initially developed for transport network problems in our paper [10], where the mathematical modeling of network problems can be represented as *a translation* between two domain specific languages (DSL1,2). One is a network description language (NET), used in the corresponding discipline (e.g., electrochemistry, gas transport, water supply, etc.), another one is a problem description language (PRO), understood by generic non-linear solvers (e.g., IPOPT, Mathematica, Matlab, etc.). The generic algorithm (Universal Translation, UT) uses a configurable set of translation rules (Translation Matrix, TM) for translating the network from one representation to another:

$$NET \times TM \to PRO. \tag{1}$$

In application to chemical kinetics, the reaction hypergraph from Table I is recorded in a symbolic form (NET):

$$\text{reaction}_i = \sum_j a^L_{ij} g_j - a^R_{ij} g_j, \tag{2}$$

where g_j are the reagents (e.g., OH^-, CH_3OH), a^L_{ij}, a^R_{ij} are the incidence matrices of the hypergraph multiplied to *stoichiometric coefficients*. These integer-valued coefficients indicate how many molecules are spent or produced, for the left and right hand side of the reaction. The translation matrix (TM) defines the rules of translation, enlisting the reagents, variables, constants, parameters, an excerpt is shown in Table II. The hypergraph is then translated to the reaction rates

$$r_i = k^L_i \prod_j (c_j)^{a^L_{ij}} - k^R_i \prod_j (c_j)^{a^R_{ij}}, \tag{3}$$

indicating how many reactions per second are happening. The rates are defined by *a probability* of the reagents to meet each other for the reaction, proportional to the product of concentrations c_j (that can be molar, volumetric, surface, etc.), in the corresponding integer powers. E.g., the reaction r_5 in Table I requires one molecule CO_{ad} (reagent 4, see Table II) to meet two molecules of OH_{ad} (reagent 1), and the corresponding reaction rate will be proportional to $c_4 c_1^2$. The proportionality coefficients k_i are important model parameters, which should be reconstructed from the experiments. Further, the reaction rates are assembled to molar balance description:

$$F_j = -\sum_i r_i (a^L_{ij} - a^R_{ij}), \tag{4}$$

indicating how many molecules (or, in appropriate normalization, moles) of the reagents are spent or produced per second. Then ordinary differential equations (ODEs) governing the chemical kinetics are formed:

$$d\nu_i/dt = F_i, \tag{5}$$

where ν_i is a molar amount for the i-th reagent.

In particular applications, these generic formulae can be modified by different normalizations, e.g., some of the reagents can be adsorbed on the electrode and are represented by dimensionless surface coverage ratios, with the range $\theta_i \in [0, 1]$. Also, a peculiarity of electrochemistry is that the electrons also belong to the reagents and their flow (electrons per second) defines the cell current measured in the experiments.

Applying these translation rules to our reaction network, we obtain [3]:

$$\begin{aligned}
&r_1 = k_1 c_1 \theta_0 - k_{-1} \theta_1, \\
&r_2 = k_2 c_2 \theta_0 - k_{-2} \theta_2, \\
&r_3 = k_3 \theta_2 \theta_1^3 - k_{-3} \theta_3 c_3^3, \\
&r_4 = k_4 \theta_3 \theta_1,\ r_5 = k_5 \theta_4 \theta_1^2, \\
&r_7 = k_7 \theta_3 \theta_1^2,\ r_8 = k_8 \theta_5, \\
&r_9 = k_9 \theta_4 \theta_1,\ r_{10} = k_{10} \theta_5 \theta_1, \\
&r_{12} = k_{12} c_1 \theta_1 - k_{-12} c_3 \theta_6,
\end{aligned} \tag{6}$$

where $\theta_0 = 1 - \sum_{i=1}^6 \theta_i$ is a part the surface of platinum catalyst, not covered by any reagent, $\theta_0 \in [0, 1]$. For the reactions, involving electrons, the Tafel equation is used:

$$\begin{aligned}
&k_1 = k^0_1 \exp(\alpha\beta\eta), \\
&k_{-1} = k^0_{-1} \exp(-(1-\alpha)\beta\eta), \\
&k_8 = k^0_8 \exp(-(1-\alpha)\beta\eta), \\
&k_{12} = k^0_{12} \exp(\alpha\beta\eta), \\
&k_{-12} = k^0_{-12} \exp(-(1-\alpha)\beta\eta), \\
&\beta = F/(RT),
\end{aligned} \tag{7}$$

where T is the absolute temperature and other constants are given in Table II,

$$\begin{aligned}
&F_1 = (r_1 - 3r_3 - r_4 - 2r_5 - 2r_7 - r_9 - r_{10} - r_{12})/C_{act}, \\
&F_2 = (r_2 - r_3)/C_{act}, \\
&F_3 = (r_3 - r_4 - r_7)/C_{act}, \\
&F_4 = (r_4 - r_5 - r_9)/C_{act}, \\
&F_5 = (r_7 - r_8 + r_9 - r_{10})/C_{act}, \\
&F_6 = r_{12}/C_{act}, \\
&F_7 = (-r_1 + r_8 - r_{12}) \cdot FA/C_{dl},
\end{aligned} \tag{8}$$

here the constants C_{act}, C_{dl}, A are also given in Table II, these 3 constants depend on the experimental setup and recalibrated each time when it is changed. The resulting ODE system looks like:

$$\begin{aligned}
&d\theta_i/dt = F_i(\theta, \eta),\ i = 1 \ldots 6, \\
&d\eta/dt = F_7(\theta, \eta) + I_{cell}/C_{dl},
\end{aligned} \tag{9}$$

where η is the cell voltage and I_{cell} is the cell current. The model is fitted to the experiment using L_2-norm of variation

$$L_2 = \left(\sum_i (I_{cell,i} - I^{exp}_{cell,i})^2 \right)^{1/2}, \tag{10}$$

with the reaction constants $k^0_i \geq 0$ in (7) and $k_i \geq 0$ for all others used as fitting parameters. The details of fitting depend on the selected experimental method.

PC method: sets all time derivatives to zero, to define a stationary point

$$\begin{aligned}
&0 = F_i(\theta^*, \eta^*),\ i = 1 \ldots 6, \\
&0 = F_7(\theta^*, \eta^*) + I^*_{cell}/C_{dl},
\end{aligned} \tag{11}$$

for fixed η^*, stepwise scanned in the range $[\eta_{min}, \eta_{max}]$, the obtained 7x7 polynomial system is solved with respect to

TABLE I. The reaction list.

r_1:	OH^-	+	Pt	$\leftrightarrow$	OH_{ad}	+	e^-		
r_2:	CH_3OH	+	Pt	$\leftrightarrow$	CH_3OH_{ad}				
r_3:	CH_3OH_{ad}	+	$3OH_{ad}$	$\leftrightarrow$	CHO_{ad}	+	$3H_2O$		
r_4:	CHO_{ad}	+	OH_{ad}	$\rightarrow$	CO_{ad}	+	H_2O		
r_5:	CO_{ad}	+	$2OH_{ad}$	$\rightarrow$	CO_2	+	H_2O	+	$2Pt$
r_7:	CHO_{ad}	+	$2OH_{ad}$	$\rightarrow$	$COOH_{ad}$	+	H_2O	+	Pt
r_8:	$COOH_{ad}$	+	e^-	$\rightarrow$	$HCOO^-$	+	Pt		
r_9:	CO_{ad}	+	OH_{ad}	$\rightarrow$	$COOH_{ad}$	+	Pt		
r_{10}:	$COOH_{ad}$	+	OH_{ad}	$\rightarrow$	CO_2	+	H_2O	+	$2Pt$
r_{12}:	OH^-	+	OH_{ad}	$\leftrightarrow$	PtO	+	H_2O	+	e^-

TABLE II. Numeration and values for the model variables and constants.

Variables		*Concentrations*		*Constants*	*Names*	*Values*
θ_1	OH_{ad}			F	Faraday constant	$9.649 \cdot 10^4$ C/mol
θ_2	CH_3OH_{ad}	c_1	OH^-	R	universal gas constant	8.314 J/(K mol)
θ_3	CHO_{ad}	c_2	CH_3OH	α	charge transfer coefficient	0.5
θ_4	CO_{ad}	c_3	H_2O	C_{dl}	capacitance	$1.899 \cdot 10^{-4}$ F
θ_5	$COOH_{ad}$			C_{act}	activity constant	$8.523 \cdot 10^{-5}$ mol/m^2
θ_6	PtO			A	electrode area	$2.376 \cdot 10^{-5}$ m^2

$(\theta^*_{1-6}, I^*_{cell})$ by *Mathematica* `NSolve` algorithm. Only real roots with $\theta^*_{0-6} \in [0,1]$ are selected. There are always 2 such roots, one is trivial: $\theta^*_4 = 1$, other $\theta^*_i = 0$, $I^*_{cell} = 0$ (electrode is completely blocked by CO_{ad}), another is non-trivial: all $\theta^*_i > 0$, $I^*_{cell} > 0$. Only non-trivial root is selected and used for minimization of L_2-norm (10).

EIS method: linearizes the system near the stationary point

$$dv/dt = Jv + b,\ J_{ij} = \partial F_i/\partial x_j, \tag{12}$$

with $n \times n$ Jacobi matrix J, n-dimensional vector of variables $x = (\theta_1, ..., \theta_{n-1}, \eta)^T$ (in our case $n = 7$) and variations of vectors $v = \delta x$, $b = (0, ..., 0, \delta I_{cell}/C_{dl})^T$. In more details we have considered this system in our previous paper [4]. The equations describe the linearized evolution around the stationary point, when the voltage is varied along the given profile. In our case, harmonic oscillations of small amplitude are considered, in complex denotation: $\delta\eta = \eta_0 \exp(i\omega t)$, $\delta I_{cell} = I_0 \exp(i\omega t)$. Their ratio gives the complex resistance, or impedance $Z = \eta_0/I_0$ of the cell. After harmonic substitution $v = v_0 \exp(i\omega t)$, $b = b_0 \exp(i\omega t)$ we have

$$(i\omega - J)v_0 = b_0 \tag{13}$$

and finally

$$C_{dl}Z(\omega) = ((i\omega - J)^{-1})_{nn}. \tag{14}$$

It can be rewritten in one of the equivalent forms:

$$C_{dl}Z(\omega) = Q_{n-1}(i\omega)/Q_n(i\omega), \tag{15}$$

where Q_k are polynomials of k-th order,

$$C_{dl}Z(\omega) = \prod_{j=1}^{n-1}(i\omega - q_j)/\prod_{j=1}^{n}(i\omega - p_j), \tag{16}$$

with poles p_j and zeros q_j of Z,

$$C_{dl}Z(\omega) = \sum_{j=1}^{n} r_j/(i\omega - p_j), \tag{17}$$

with residues r_j at the corresponding poles p_j. Note that p_j are eigenvalues of Jacobi matrix J, while q_j are eigenvalues of its left-upper $(n-1) \times (n-1)$ submatrix.

The fitting involves a solution of Non-Linear Program (NLP) of the form

$$\text{find } \min_x f(x),\ \text{such that } g(x) = 0 \text{ and } h(x) \geq 0, \tag{18}$$

where the equality conditions g include the equations for stationary point (11), the definition of Jacobi matrix in (12) and the definition of poles and zeros

$$Q_{n-1}(q_j) = 0,\ Q_n(p_j) = 0, \tag{19}$$

the inequality conditions h include the already mentioned

$$k_r \geq 0,\ \theta^*_i \geq 0,\ \theta^*_0 = 1 - \sum \theta^*_i \geq 0. \tag{20}$$

In the case, if the obtained system becomes overdetermined, some of the equations can be moved to the target function, e.g.,

$$f(x):\ L_2 = \textstyle\sum_j |Q_{n-1}(q_j)|^2 + \sum_j |Q_n(p_j)|^2. \tag{21}$$

The solution is performed by *Mathematica* `NMinimize` algorithm. The typical model shapes on the Nyquist plot corresponding to a different number and types of poles and zeros are shown on Figure 4. One of these forms has been used on Figure 5 for testing of the reconstruction algorithm. The paper [4] provides further details on the structure of solution. In particular, if some poles and zeros come too close to each other, they should be mutually cancelled to increase stability of the reconstruction. The stability condition has a

form $N_{exp}(N_p + N_z + 1) \geq N_k$, where N_{exp} is the number of experiments, $N_{p,z}$ is the number of reconstructed poles and zeros per experiment, N_k is the number of reaction constants. The errors of the reconstructed Nyquist plot can be controlled by the formula

$$\delta Z(\omega)/Z(\omega) = -\sum \delta q_j/(i\omega - q_j) + \sum \delta p_j/(i\omega - p_j). \quad (22)$$

CV method: considers the original ODE system (9). We solve this system directly via numerical integration by *Mathematica* `NDSolve` algorithm. The starting point of the integration should be selected to provide cyclicity of solution $\theta(T_p) = \theta(0)$, T_p is a period. Alternatively, the system should be integrated during several (3-5) "warming up" periods till the cycle becomes reproduced. The resulting I_{cell} is used for fitting of L_2-norm (10) by *Mathematica* `NMinimize` algorithm. The main problem is a determination of a region in multi-dimensional ($dim = 14$ in our case) space of the fitting parameters, the reaction coefficients $k_i^{(0)}$, where the starting point for fitting procedure is located, close enough to the minimum searched. We solved this problem in [2] by an iterative direct search Monte Carlo method in a combination with interactive visualization. Figure 6 shows the visualization tool, implemented by means of *Mathematica* `Manipulate` algorithm. The user can interactively change the reaction constants and see the obtained integrated evolution in comparison with the experimental data.

The first interesting observation, obtained by the described method, is that CV diagrams at low voltages contain a PC part, where the shapes for increasing and decreasing voltage coincide and the time derivatives in (9) can be omitted. These parts, displayed on Figure 7 for the experiments with different volumetric concentrations of the reagents, can be fitted by the stationary curves from the above described PC method. Although this method does not require computationally expensive numerical integration and is very fast, its disadvantage is that only the ratios of reaction coefficients can be reconstructed. Indeed, for the absent time derivatives in (11), every equation can be divided to one of the reaction constants, preserving its validity.

At larger voltages, *a hysteresis* effect appears, see Figure 8. It is purely dynamic effect, related with the presence of time derivative in (9), which after time-reflection reverses its sign. As a result, the shapes of volt-ampere characteristics for increasing and decreasing voltage do not coincide. Our analysis in [2], [3] shows that the hysteresis effect appears when some intermediates in the reaction network are weakly coupled. In this case the corresponding θ evolution becomes retarded with respect to the voltage variation. We have tried to decouple several intermediates, indicated by orange bubbles in Figure 3, by strongly reducing their reaction constants. The best results are obtained with decoupling of the 6th reagent, PtO.

Next, Figure 9 shows CV plots in experiments with different upper voltage η_{max}. The data are described in details in [12], where the focus is on the physical-chemical processes, while in the current paper we focus on the simulation methodology and the parameter identification of the mathematical model. It is visible in Figure 9, how hysteresis is reduced and finally disappears, leaving PC behavior only, when η_{max} is decreasing. Again, the hysteresis effect vanishes when η_{max} is shifted below the voltage region, where the production of PtO happens. Therefore, the PC and EIS methods should be preferably applied in this low voltage region.

The experimental data on Figures 7 and 8 are well fitted by the model. This fit is performed individually for every plot, presenting given values of concentrations, as described in [3]. An attempt to fit all experiments by a single set of reaction constants fails, which has been interpreted in [3] as a dependence of reaction constants on concentrations.

The experimental data on Figure 9 correspond to the same concentration values and variable η_{max}. Three columns correspond to three isolated optima of the fit, presented on Figure 10. The best fit is provided by the left column, set 1. It is visible that the upper increasing curve always has a better fit than the lower decreasing one. It is also visible that the shape of the increasing curve is the same for all experiments, only the upper limit is different. This happens because we are fitting the first cycle of CV plot. Comparing the plots in different rows, we can conclude that the system initially "does not know", when its η will be reverted, and produces the same curve upto this moment. We also observe that the decreasing curve is sensitive to the small values of θ_0 in the region of upper voltage and fluctuates when the model parameters are slightly changed. More uncertainty is related with the unknown starting θ-values for the evolution on the first cycle, for which in our experiments the clean electrode $\theta_0 = 1$ was assumed. There is also a large visible scatter in the data, corresponding to different cycles of the CV plot. Like it happens with slow approaching of the equilibrium in PC experiments, here we observe a slow approaching of the limit cycle in CV plots.

IV. Improvements of CV method

In this section, we consider a possibility that the ODE system (9) can be reduced to the system of Differential-Algebraic Equations (DAE), where some of the equations keep time derivative terms and the others do not. Let us write the equations in the form:

$$\alpha_i d\theta_i/dt = F_i = \textstyle\sum_j C_{ij} r_j, \quad (23)$$

where C_{ij} is a structural matrix relating production rates F_i and reaction rates r_j and α_i are constant coefficients, for the case of ODE set to $\alpha_i = 1$ and for the case of DAE to $\alpha_i = 0$. We have also reassigned normalization factors between F_i and r_j, so that both are measured in the same units (s^{-1}). The measured quantity is a cell current, given by the expression:

$$I_{cell} = FAC_{act}F_7, \; F_7 = \textstyle\sum_j C_{7j} r_j. \quad (24)$$

Here, we add the 7th row in the structural matrix and omit the practically vanishing capacitance term $C_{dl}d\eta/dt$.

Now, we draw attention to Figure 11, which depicts the evolution of production and reaction rates. It is visible that some r_j compensate each other, resulting in almost zero F_i. This common property, also noted in [11], means that some of the reactions proceed so fast that they are almost permanently in equilibrium. One production rate is not in equilibrium. It is also characterized by the presence of only one reaction: $F_6 = r_{12}$. Thus, in the equations, one can switch off the dynamic terms for all reagents except for the 6th, so that $\alpha_i = \delta_{i6}$. As a result, the ODE system is replaced by an equivalent DAE system. Mathematica v11 can be used to solve DAE systems with the same efficiency as ODE.

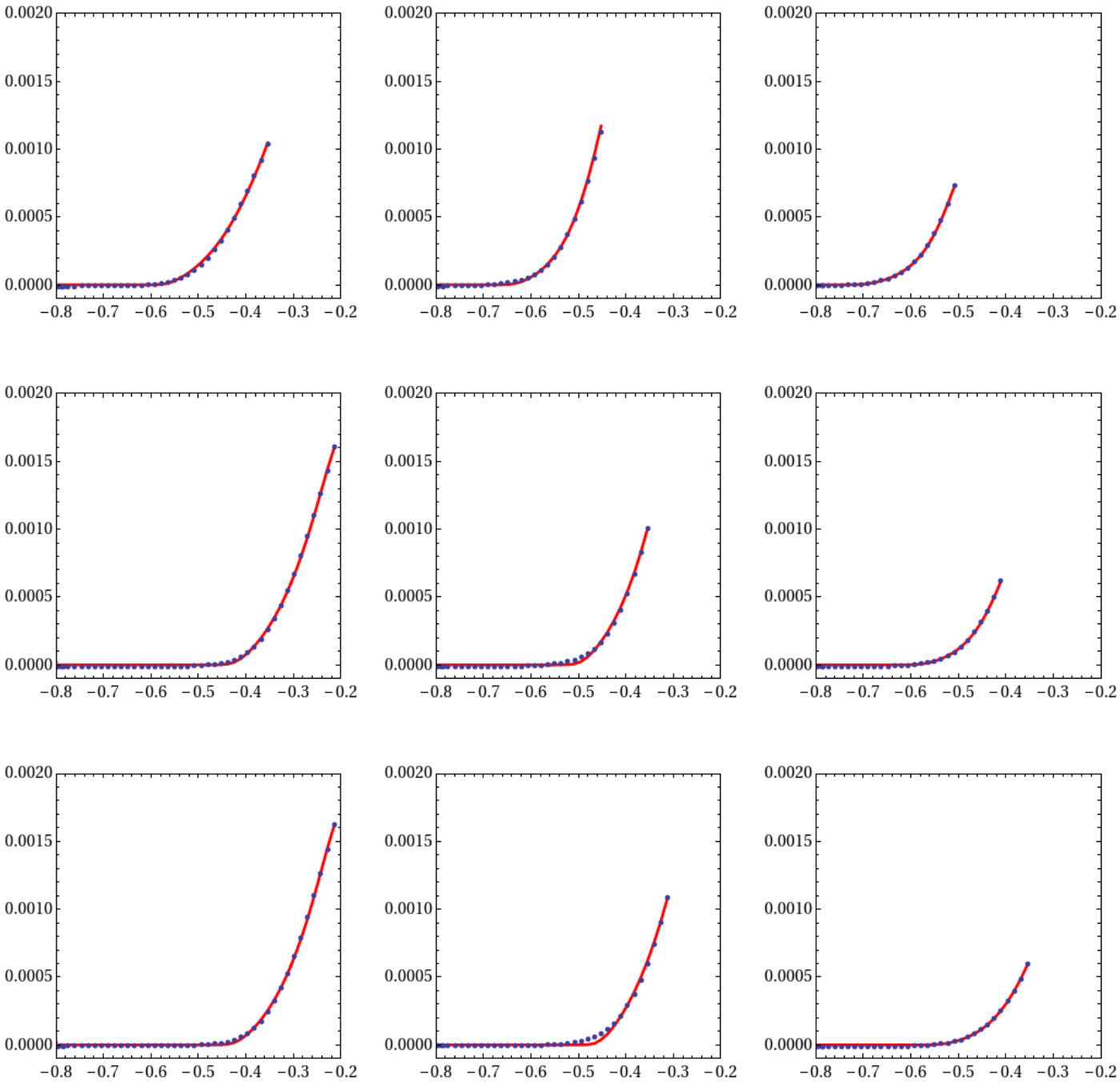


Figure 7. PC experiments. From left to right, the concentration of alkaline KOH is set to $c_1 = \{0.1, 0.5, 1.0\}M$, from bottom to top, the concentration of methanol CH_3OH is set to $c_2 = \{0.5, 0.75, 1.0\}M$. Horizontal axes represent the voltage in Volts, vertical axes – the cell current in Amperes. Blue points are experimental data. Red lines show the best fit by the model with reaction constants reconstructed separately for each experiment.

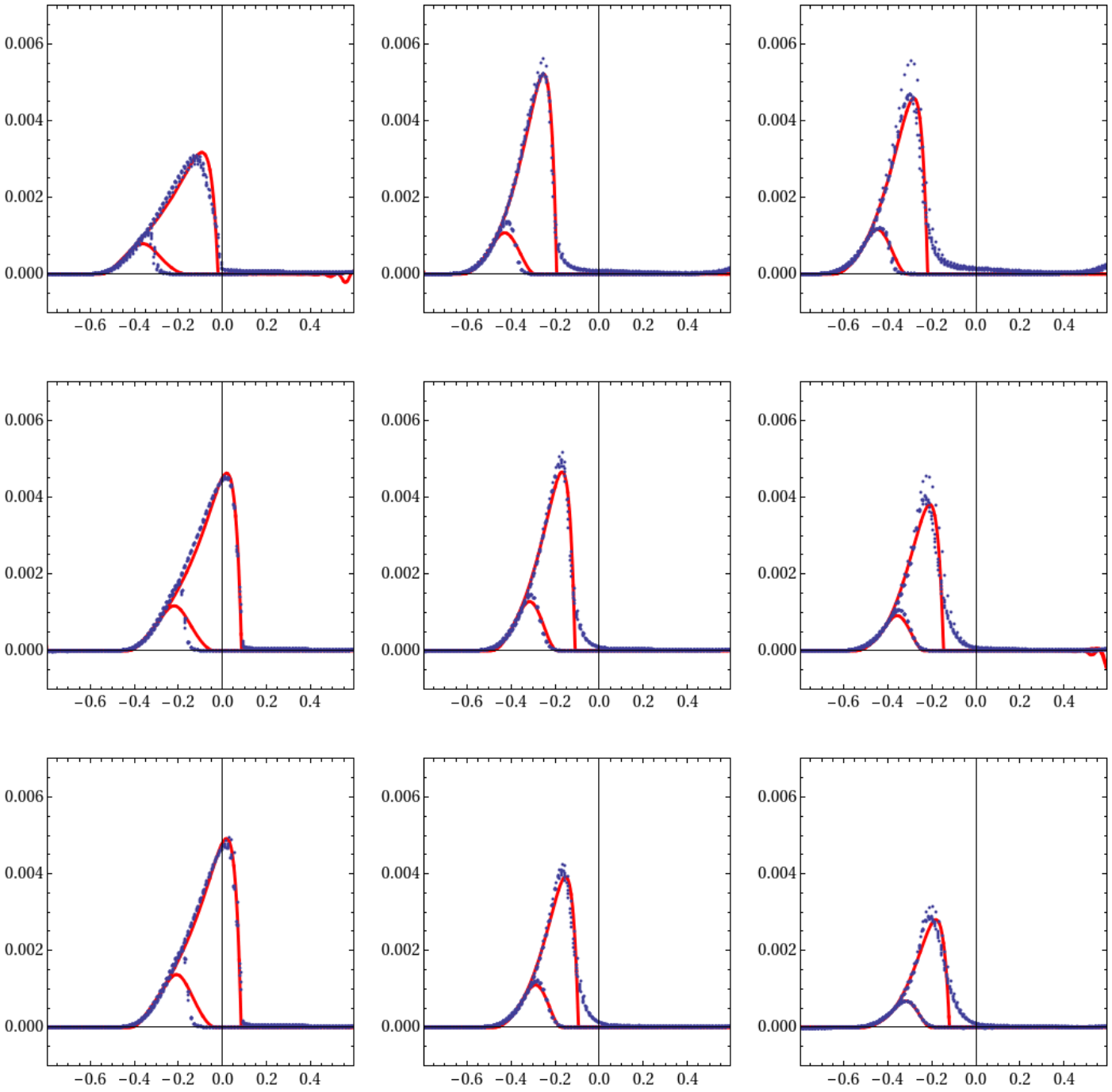


Figure 8. CV experiments. From left to right, the concentration of alkaline KOH is set to $c_1 = \{0.1, 0.5, 1.0\}M$, from bottom to top, the concentration of methanol CH_3OH is set to $c_2 = \{0.5, 0.75, 1.0\}M$. Horizontal axes represent the voltage in Volts, vertical axes – the cell current in Amperes. Blue points are experimental data. Red lines show the best fit by the model with reaction constants reconstructed separately for each experiment.

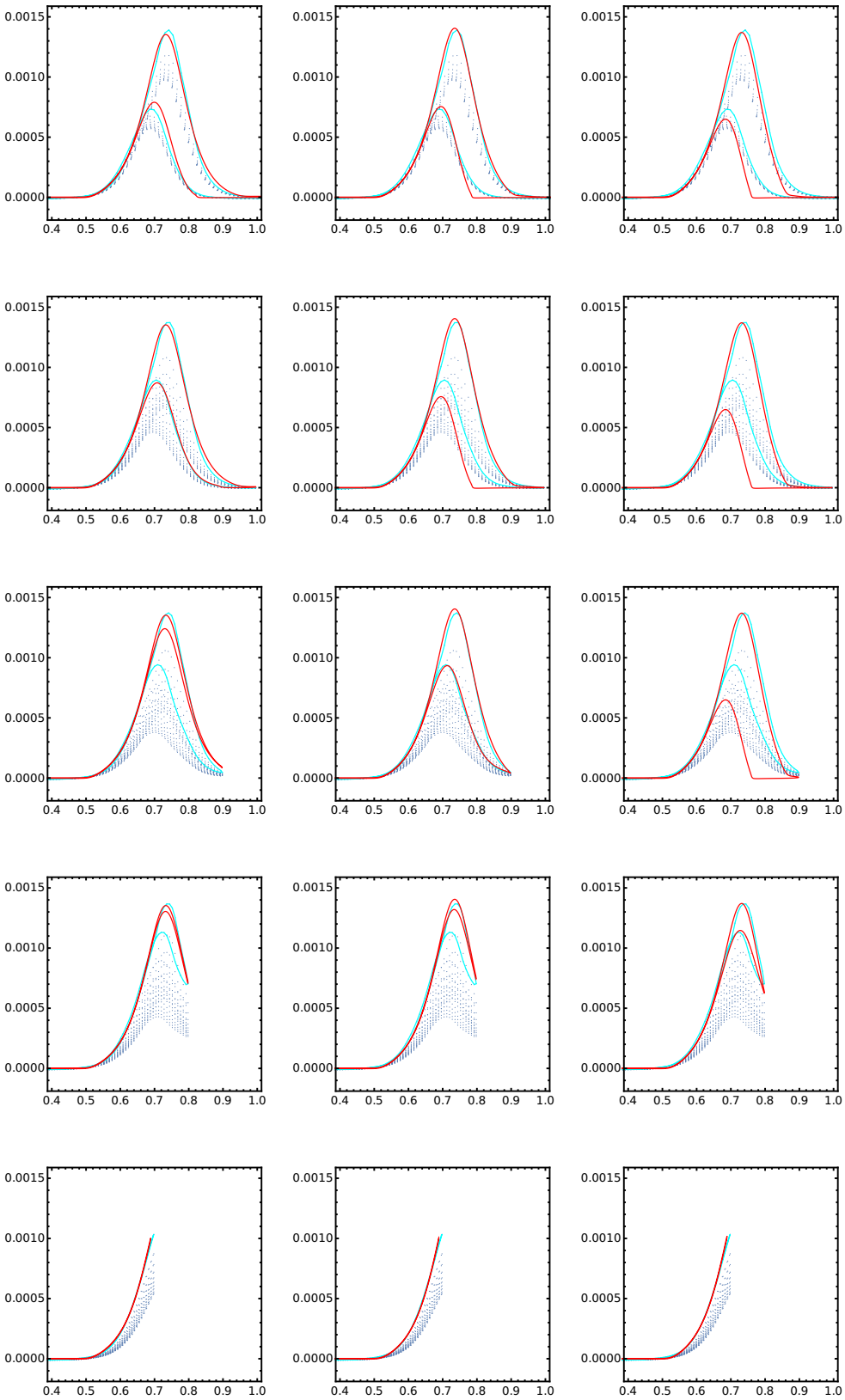


Figure 9. CV experiments, with different upper voltage. Three columns correspond to three isolated optima of the fit. From top to bottom in every column – upper voltage is reduced. Horizontal axes represent the voltage in Volts, vertical axes – the cell current in Amperes. Blue points are experimental data for all cycles. Cyan lines are experimental data for the first cycle. Red lines show the best fit of the first cycle by the model. Data from [12].

Log Constants	*sets*		
	1	2	3
p_1^0	-8.52902	-8.54902	-8.52902
p_{-1}^0	1.23059	1.37059	1.38059
p_2	1.75957	1.89957	1.90957
p_{-2}	7.82156	7.96156	7.97156
p_3	8.2795	8.4195	8.4295
p_{-3}	0.899914	1.03991	1.04992
p_4	-0.94845	-0.808451	-1.14845
p_5	-7.56233	-7.42233	-7.41233
p_7	-4.38496	-4.24496	-4.23496
p_8^0	2.87557	2.89557	2.88557
p_9	-1.9024	-1.7624	-1.75239
p_{10}	-7.86652	-7.72651	-7.71652
p_{12}^0	-13.0262	-12.1362	-11.5962
p_{-12}^0	1.04633	1.07133	1.00133

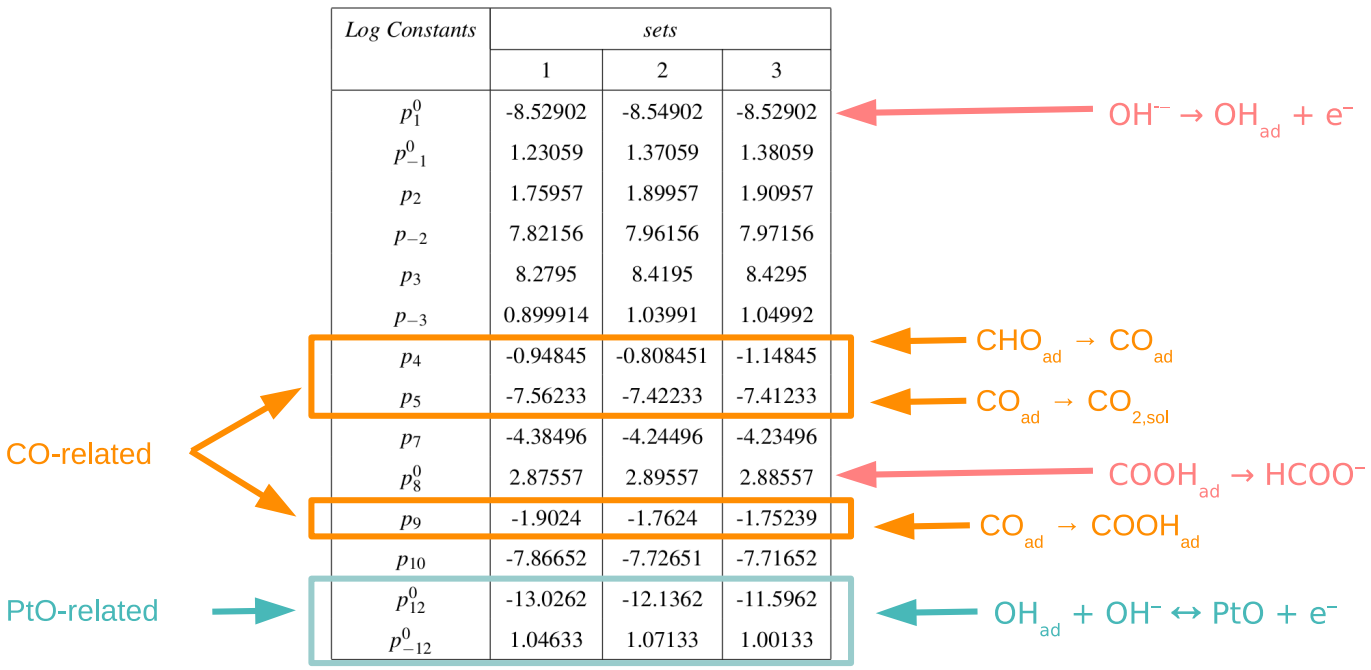


Figure 10. Reconstructed reaction constants for the experiments with different upper voltage. The constants are given in logarithmic values $p_i = \log_{10}(k_i/[mol/(m^2 s)])$. Two sources of decoupling of reagents (CO-related, PtO-related) are indicated, with corresponding reactions.

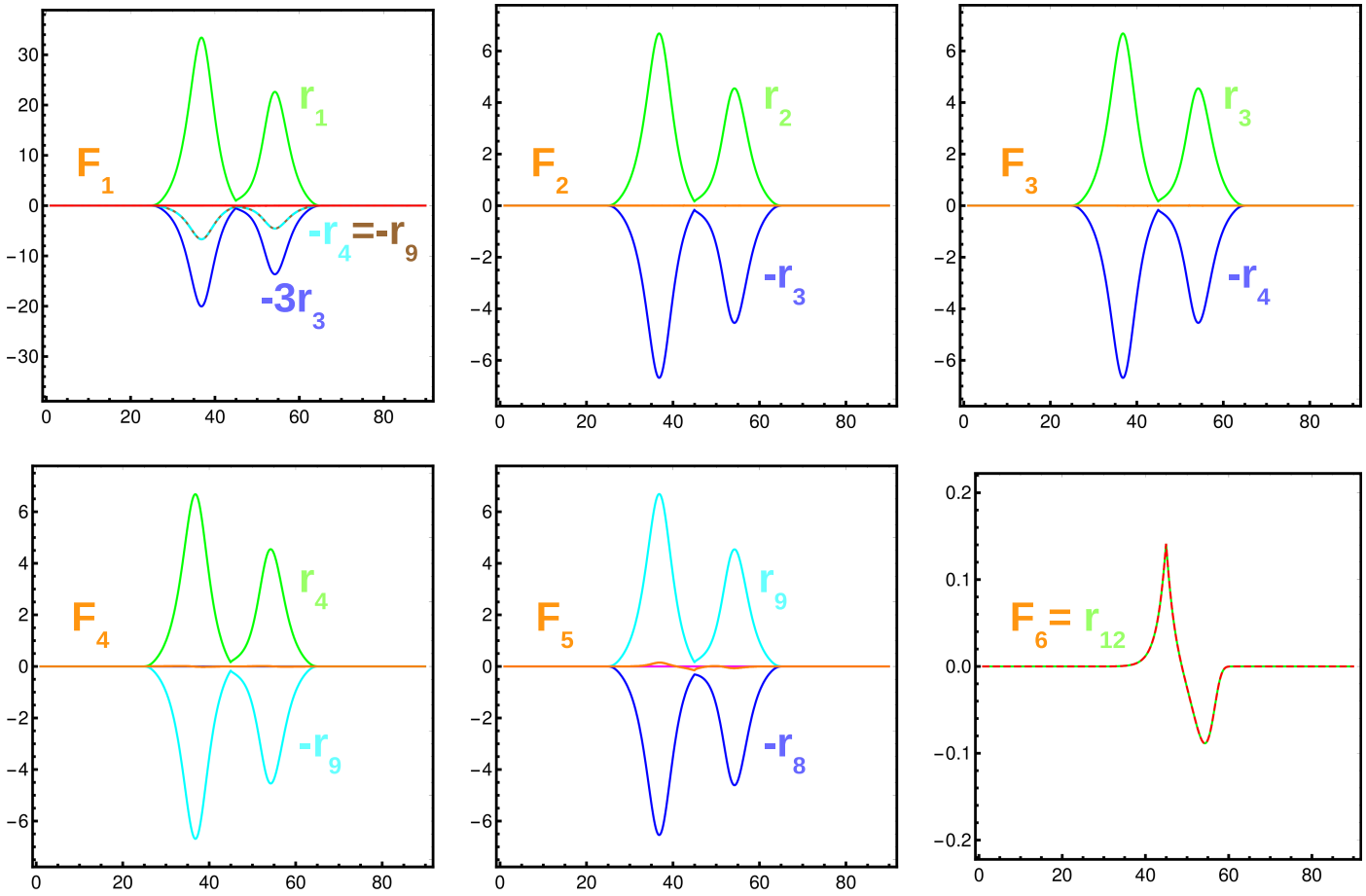


Figure 11. The plots of production rates F_i and reaction rates r_i. All production rates except F_6 show an approach to equilibrium. The horizontal axes show the time in seconds, the vertical axes: F_i and r_i in s^{-1}.

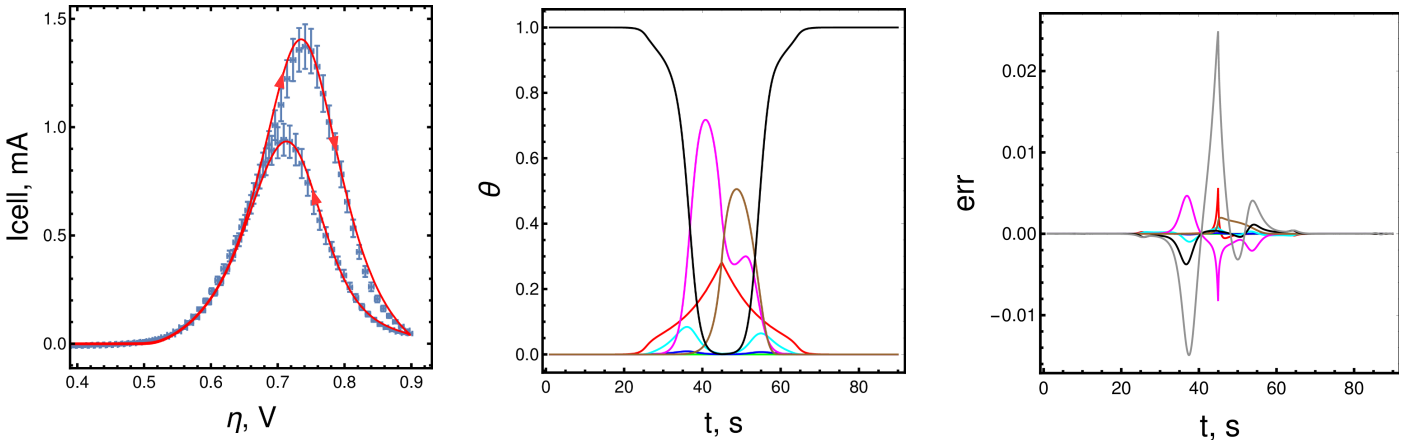


Figure 12. On the left: CV plot, blue points with error bars – the experiment, red line – the model. In the center: evolution of surface coverages, the colors (red, green, blue, cyan, magenta, brown) encode sequential θ_i, black shows the free Pt surface. On the right: ODE→DAE variations for θ_i (the same colors), relative variation for I_{cell} (in gray).

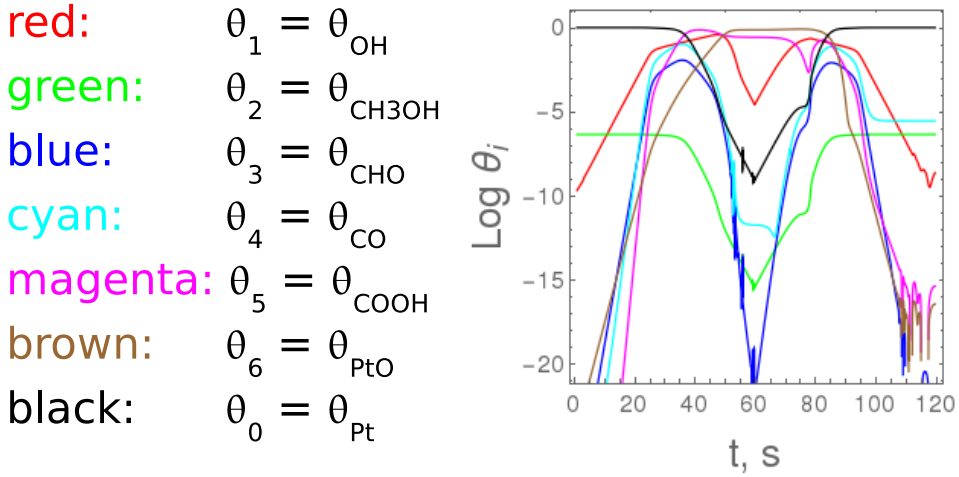


Figure 13. Evolution of θ_i in logarithmic values.

The described mathematical model fits the experimental data well, both for ODE and for DAE formulations, see Figure 12 left. After the transition to DAE, the CV plot in Figure 12 left changes slightly, as well as the detailed evolution of θ_i, shown in Figure 12 center and in logarithmic values in Figure 13. An interesting property that immediately catches the eye is the temporal asymmetry of the profiles for some reagents. Since the voltage is an even periodic function, if all reactions were in equilibrium, all θ_i would be even periodic. They would behave like red or black lines, corresponding to OH_{ad} and free Pt in Figure 12. Deviation from this behavior for magenta and brown, that is, $COOH_{ad}$ and PtO, is a purely dynamic effect. The consequence of this effect is the observed mismatch (hysteresis) for the increasing and decreasing branches of the CV plot. In line with this work, it is important that DAE provides essentially the same profiles as ODE. Figure 12 right measures the deviation between the DAE and ODE, for θ_i, in the same colors, as well as the deviation of I_{cell} relative to its maximum, shown in gray. As a result, the transition from ODE to DAE results in 0.8% maximal variation for θ_i and 2.5% for I_{cell}, proving a good accuracy of the DAE representation.

V. Conclusion

In this paper, alkaline methanol oxidation has been considered, an important electrochemical process in the design of efficient fuel cells. The reaction network, represented as a hypergraph of reactions, connecting multiple reagents, is automatically translated to the mathematical model, including an ODE system describing the kinetics of the process. The difference between the modeled and experimentally measured current of the cell is used for the fitting of the parameters of the underlying mathematical model. Three types of experiments (PC, EIS, CV) can be used for the identification of the parameters.

Further, the model reduction can be performed by setting fast reactions to the equilibrium and leaving the dynamical term only for slow reactions. The obtained DAE formulation has an advantage that only one degree of freedom (surface

coverage by PtO) remains in the system, to which the evolution of other reagents is strictly coupled. The model is reduced and still describes the same effects as the complete system. In particular, it explains the dynamic hysteresis of volt-ampere characteristics of the cell. The methods have been tested on a range of experiments, including different concentrations of the reagents and different voltage range.

VI. Acknowledgment

The work has been partially supported by the German Federal Ministry for Economic Affairs and Energy, grant BMWI-0324019A, project MathEnergy and by the German Bundesland North Rhine-Westphalia, the European Regional Development Fund, grant Nr. EFRE-0800063, project ES-FLEX-INFRA.

References

[1] T. Clees, B. Klaassen, I. Nikitin, L. Nikitina, S. Pott, U. Krewer, and T. Haisch, "Model Reduction in the Design of Alkaline Methanol Fuel Cells", Proc. of INFOCOMP 2019, July 28, 2019 to August 02, 2019 - Nice, France, pp. 21-22, Pub. IARIA 2019, ISBN: 978-1-61208-732-0.

[2] T. Clees, I. Nikitin, L. Nikitina, S. Pott, U. Krewer, and T. Haisch, "Parameter Identification in Cyclic Voltammetry of Alkaline Methanol Oxidation", in Proc. SIMULTECH 2018, July 29-31, 2018, Porto, Portugal, pp. 279-288, ISBN: 978-989-758-323-0.

[3] T. Clees, I. Nikitin, L. Nikitina, S. Pott, U. Krewer, and T. Haisch, "Mathematical Modeling of Alkaline Methanol Oxidation for Design of Efficient Fuel Cells", in: Obaidat M., Ören T., Rango F. (eds) Simulation and Modeling Methodologies, Technologies and Applications, SIMULTECH 2018, Advances in Intelligent Systems and Computing, vol 947, 2020 (First Online 20 November 2019), pp. 181-195, Springer, DOI: 10.1007/978-3-030-35944-7_9.

[4] T. Clees, I. Nikitin, L. Nikitina, D. Steffes-lai, S. Pott, U. Krewer, and T. Windorfer, "Electrochemical Impedance Spectroscopy of Alkaline Methanol Oxidation", in Proc. INFOCOMP 2017, The Seventh International Conference on Advanced Communications and Computation, pp. 46-51, IARIA, 2017.

[5] U. Krewer, T. Vidakovic-Koch, and L. Rihko-Struckmann, "Electrochemical oxidation of carbon-containing fuels and their dynamics in low-temperature fuel cells", ChemPhysChem, vol. 12, 2011, pp. 2518-2544.

[6] U. Krewer, M. Christov, T. Vidakovic, and K. Sundmacher, "Impedance spectroscopic analysis of the electrochemical methanol oxidation kinetics", Journal of Electroanalytical Chemistry, vol. 589, 2006, pp. 148-159.

[7] B. Beden, F. Kardigan, C. Lamy, and J. M. Leger, "Oxidation of methanol on a platinum electrode in alkaline medium: effect of metal ad-atoms on the electrocatalytic activity", J. Electroanalytical Chem., vol. 142, 1982, pp. 171-190.

[8] F. Ciucci, "Revisiting parameter identification in electrochemical impedance spectroscopy: Weighted least squares and optimal experimental design", Electrochimica Acta, vol. 87, 2013, pp. 532-545.

[9] A. J. Bard and L. R. Faulkner, Electrochemical Methods: Fundamentals and Applications, Wiley 2000, ISBN: 978-0-471-04372-0.

[10] A. Baldin et al., "Universal Translation Algorithm for Formulation of Transport Network Problems", in Proc. SIMULTECH 2018, vol. 1, pp. 315-322.

[11] A. N. Gorban, "Model reduction in chemical dynamics: slow invariant manifolds, singular perturbations, thermodynamic estimates, and analysis of reaction graph", Current Opinion in Chemical Engineering, vol. 21, 2018, pp. 48-59, DOI: 10.1016/j.coche.2018.02.009.

[12] T. Haisch et al., "The origin of the hysteresis in cyclic voltammetric response of alkaline methanol electrooxidation", submitted to Physical Chemistry Chemical Physics.

Multi-Agents Spatial Visibility Trajectory Planning and Patrolling Using Inverse Reinforcement Learning

Oren Gal and Yerach Doytsher
Mapping and Geo-information Engineering
Technion - Israel Institute of Technology
Haifa, Israel
e-mails: {orengal, doytsher}@technion.ac.il

***Abstract*—In this paper, we present a conceptual Spatial Trajectory Planning (STP) method using Rapid Random Trees (RRT) planner, generating visibility motion primitives in urban environments using Inverse Reinforcement Learning (IRL) approach. Visibility motion primitives are set by using Spatial Visibility Clustering (SVC) analysis. Based on the STP planning method, we introduce IRL formulation and analysis which learns the value function of the planner from demonstrated trajectories and generating spatial visibility trajectory planning. Additionally, we study the visible trajectories planning for patrolling application using heterogeneous multi agents in 3D urban environments. Our concept is based on spatial clustering method using visibility analysis of the 3D visibility problem from a viewpoints in 3D urban environments, defined as locations. We consider two kinds of agents, with different kinematic and perception capabilities. Using simplified version of Traveling Salesman Problem (TSP), we formulate the problem as patrolling strategy one, with upper bound optimal performances. We present a combination of relative deadline UniPartition approaches based on visibility clusters. These key features allow new planning optimal patrolling strategy for heterogeneous agents in urban environment. We demonstrate our patrolling strategy method in simulations using Autonomous Navigation and Virtual Environment Laboratory (ANVEL) test bed environment.**

Keywords-Visibility; 3D; Spatial analysis; Motion Planning.

I. INTRODUCTION AND RELATED WORK

Spatial clustering in urban environments is a new spatial field from trajectory planning aspects (Gal and Doytsher 2014). The motion and trajectory planning fields have been extensively studied over the last two decades (Bellingham et al. 2002; Bortoff 2000; Chitsaz and LaValle 2007; Erdmann and Lozano-Perez 1987; Fiorini and Shiller 1998; Fraichard 1999; Latombe 1990; LaValle 1998; LaValle 2006; LaValle and Kuffner 1999; Sasiadek and Duleba 2000).

The main effort has focused on finding a collision-free path in static or dynamic environments, i.e., in moving or static obstacles, using roadmap, cell decomposition, and potential field methods (Gal and Doytsher 2013; Obermeyer 2009; Shaferman and Shima 2008).

The path-planning problem becomes an NP-hard one, even for simple cases such as time-optimal trajectories for a system with point-mass dynamics and bounded velocity and acceleration with polyhedral obstacles (Donald et al. 1993).

Efficient solutions for an approximated problem were investigated (LaValle and Kuffner 1999), addressing non-holonomic constraints by using the Rapidly Random Trees (RRT) method (LaValle 1998). Over the years, many other semi-randomized methods were proposed, using evolutionary programming (Capozzi and J. Vagners 2001; Lum et al. 2006; Pongpunwattana and Rysdyk 2004).

The randomized sampling algorithms planner, such as RRT, explores the action space stochastically. The RRT algorithm is probabilistically complete, but not asymptotically optimal (Karaman and Frazzoli 2011). The RRT* planner Karaman et al. 2011) challenges optimality by a rewiring process each time a node is added to the tree. However, in cluttered environments, RRT* may behave poorly since it spends too much time deciding whether to rewire or not.

Overall, only a few works have focused on spatial analysis characters integrated into trajectory planning methods such as visibility analysis or spatial clustering methods (Gal and Doytsher 2013; Shaferman and Shima 2008).

Our research contributes to the spatial data clustering field, where, as far as we know, visibility analysis has become a leading factor for the first time. The SVC method, while mining the real pedestrians' mobility datasets, enables by a visibility analysis to set the number of clusters.

The efficient computation of visible surfaces and volumes in 3D environments is not a trivial task. The visibility problem has been extensively studied over the last twenty years, due to the importance of visibility in GIS and Geomatics, computer graphics and computer vision, and robotics. Accurate visibility computation in 3D environments is a very complicated task demanding a high computational effort, which could hardly have been done in a very short time using traditional well-known visibility methods (Plantinga and Dyer 1990).

The exact visibility methods are highly complex, and cannot be used for fast applications due to their long computation time. Previous research in visibility computation has been devoted to open environments using DEM models, representing raster data in 2.5D (Polyhedral model), and do not address, or suggest solutions for, dense built-up areas.

Most of these works have focused on approximate visibility computation, enabling fast results using interpolations of visibility values between points, calculating point visibility with the Line of Sight (LOS) method (Doytsher and Shmutter 1994; Durand 1999). Lately, fast and accurate visibility analysis computation in 3D environments has been presented (Gal and Doytsher 2012; Gal and Doytsher 2013).

Multi-agents decision making and control methods can be divided into two major disciplines, centralized and decentralized approaches. The basic idea of centralized approach is to make all the decisions in one place. All tasks are concentrated by a single entity, named 'Central Task Planner and Scheduler' (CTPS).

The CTPS translates the tasks into smaller tasks (sub-tasks), which will later be sent to the appropriate agents, according to their capabilities, their assignment and their workload. Theoretically, the centralized approach appears to do the trick. It allows knowing in advance all the tasks to be done and the connections among them, allows choosing the most fitting disassembling of the problem to sub-tasks. Indeed, this is a significant advantage, as there is no disassembling which will be ideal for all missions.

However, this approach does not fit a dynamic environment, in which unpredictable events may occur. Multi-agents in marine environment usually not in a constant contact with CTPS nor with each other, even though the CTPS requires a continuous stream of data about the forthcoming events in order to provide an effective response. Solutions to this problem (such as placing multiple sensors in the environment) are expensive and hard to apply.

On the other hand, at the decentralized approach, each agent is responsible for a group of tasks, and there is no need using entity such as CTPS. A predetermined disassembling is applied on the problem, and the agents can try to contact each other, in order to improve it. As mentioned above, this solution is problematic, as there is no disassembling which will be ideal for all problems.

Despite this fact, the lack of the CTPS allows every agent to process the data it collects by itself, and, for example, plan its own trajectory using local sensors data and decide what the next action is. The benefit of this approach is, of course, the speed of reaction and the independence of the agents. Moreover, it allows real time reaction to dynamic changes in the environment. As said, this is a problematic matter in the centralized approach.

In this paper, we present, for the first time as far as know, a unique conceptual Spatial Trajectory Planning (STP) method based on RRT planner. The generated trajectories are based on visibility motion primitives set by SVC Optimal Control Points (OCP) as part of the planned trajectory, which takes into account exact 3D visible volumes analysis clustering in urban environments.

The proposed planner includes obstacle avoidance capabilities, satisfying dynamics' and kinematics' agent model constraints in 3D environments, guaranteeing probabilistic completeness. The generated trajectories are dynamic ones and are regularly updated during daylight hours due to SVC OCP during daylight hours. STP trajectories can be used for tourism and entertainment applications or for homeland security needs.

In the following sections, we first introduce the RRT planner and our extension for a spatial analysis case, such as 3D visibility. Later on, we present the STP planner, using RRT and SVC capabilities. In the last part of the paper, we present the Inverse Reinforcement Learning (IRL) approach and algorithm based on the proposed STP planning method, learning the value function of the planner from demonstrated trajectories.

II. Spatial Rapid Random Trees

In this section, the RRT path planning technique is briefly introduced with spatial extension. RRT can also deal with high-dimensional spaces by taking into account dynamic and static obstacles including dynamic and non-holonomic robots' constraints.

The main idea is to explore a portion of the space using sampling points in space, by incrementally adding new randomly selected nodes to the current tree's nodes.

RRTs have an (implicit) Voronoi bias that steers them towards yet unexplored regions of the space. However, in case of kinodynamic systems, the imperfection of the underlying metric can compromise such behavior. Typically, the metric relies on the Euclidean distance between points, which does not necessarily reflect the true cost-to-go between states. Finding a good metric is known to be a difficult problem. Simple heuristics can be designed to improve the choice of the tree state to be expanded and to improve the input selection mechanism without redefining a specific metric.

A. RRT Stages

The RRT method is a randomized one, typically growing a tree search from the initial configuration to the goal, exploring the search space. These kinds of algorithms consist of three major steps:

1. **Node Selection:** An existing node on the tree is chosen as a location from which to extend a new branch. Selection of the existing node is based on probabilistic criteria such as metric distance.
2. **Node Expansion:** Local planning applied a generating feasible motion primitive from the current node to the next selected local goal node, which can be defined by a variety of characters.

3. **Evaluation:** The possible new branch is evaluated based on cost function criteria and feasible connectivity to existing branches.

These steps are iteratively repeated, commonly until the planner finds feasible trajectory from start to goal configurations, or other convergence criteria.

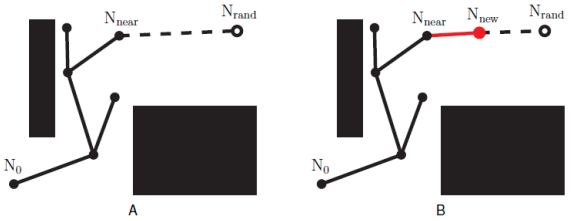


Figure 1. The RRT algorithm: (A) Sampling and node selection steps; (B) Expansion step.

A simple case demonstrating the RRT process is shown in Figure 1. The sampling step selects N_{rand}, and the node selection step chooses the closest node, N_{near}, as shown in Figure 1.A. The expansion step, creating a new branch to a new configuration, N_{new}, is shown in Figure 1.B. An example for growing RRT algorithm is shown in Figure 2.

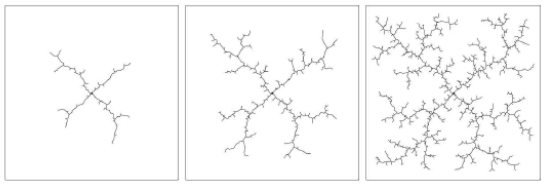


Figure 2. Example for growing RRT algorithm.

B. Spatial RRT Formulation

We formulate the RRT planner and revise the basic RRT planner for a 3D spatial analysis case for a continuous path from initial state x_{init} to goal state x_{goal}:

1. **State Space:** A topological space, *X*.
2. **Boundary Values:** $x_{init} \subset X$ and $x_{goal} \subset X$.
3. **Free Space**: A function $D: X \rightarrow \{true, false\}$ that determines whether $x(t) \subset X_{free}$ where X_{free} consist of the attainable states outside the obstacles in a 3D environment.
4. **Inputs:** A set, *U*, contains the complete set of attainable control efforts u_i, that can affect the state.
5. **Incremental Simulator:** Given a current state, $x(t)$, and input over time interval Δt, compute $x(t+\Delta t)$.
6. **3D Spatial Analysis:** A real value function, $f(x; u, OCP_i)$ which specifies the cost to the center of 3D visibility volumes cluster points (OCP) between a pair of points in X .

C. Spatial RRT Formulation

We present a revised RRT pseudo code described in Table I, for spatial case generating trajectory *T*, applying *K* steps from initial state x_{init}. The *f* function defines the dynamic model and kinematic constraints, $\dot{x} = f(x; u, OCPi)$, where *u* is the input and OCP_i set the next new state and the feasibility of following the next spatial visibility clustering point.

TABLE I. SPATIAL RRT PSEUDO CODE

```
Generate Spatial RRT (x_init; K; Δt)
T.init (x_init);
For k = 1 to K do
        x_rand ← random.state();
        x_near ← nearest.neighbor (x_rand; T );
        u ← select.input (x_rand; x_near);
        x_new ← new.state (x_near; u; Δt; f);
        T.add.vertex (x_new);
        T.add.edge (x_near; x_new; u);
End
Return T
```

III. SPATIAL TRAJECTORY PLANNING (STP)

In this section, we present a conceptual STP method based on RRT planner. The method generates visibility motion primitives in urban environments. The STP method is based on a RRT planner extending the stochastic search to specific *OCP*. These primitives connecting between nodes through OCP are defined as visibility primitives.

A common RRT planner is based on greedy approximation to a minimum spanning tree, without considering either path lengths from the initial state or following or getting close to specific OCP. Our STP planner consist of a tree's extension for the next time step with probability to goal and probability to waypoint, where trajectories can be set to follow adjacent points or through OCP. The planner includes obstacle avoidance capabilities, satisfying dynamics' and kinematics' agent model constraints in 3D environments. As we demonstrated in the previous section, the OCP are dynamic during daylight hours. Due to *OCP's* dynamic character, the generated trajectory is also a dynamic one during daylight hours.

We present our concept addressing the STP method formulating planner for a UGV model, integrating *OCP's* as part of the generated trajectories along with obstacle avoidance capability.

A. Dynamic Model

In this section, we suggest an Unmanned Ground Vehicle (UGV) dynamic model based on the four-wheeled car system (UGV) with rear-wheel drive and front-wheel steering (Lewis 2006). This model assumes that only the front wheels are capable of turning and the back wheels must

roll without slipping, and all the wheels turn around the same point (rotation center) which is co-linear with the rear axle of the car, as can be seen in Figure 19, where L is the length of the car between the front and rear axles. r_t is the instantaneous turning radius.

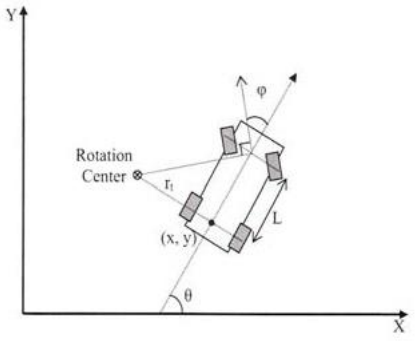


Figure 3. Four-Wheeled Car Model with Front-Wheel Steering (Lewis 2006)

Thus, UGV dynamic model can be described as:

$$\dot{x} = f(x,u) = \begin{Bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{Bmatrix} = \begin{Bmatrix} v\cos(\theta) \\ v\sin(\theta) \\ \frac{v}{L}\tan(\emptyset) \end{Bmatrix} \quad (1)$$

The state vector, x, is composed of two position variables (x,y) and an orientation variable, θ. The x-y position of the car is measured at the center point of the rear axle. The control vector, u, consists of the vehicle's velocity, v, and the angle of the front wheels, ϕ, with respect to the car's heading.

B. Search Method

Our search is guided by following spatial clustering points based on 3D visible volumes analysis in 3D urban environments, i.e., Optimal Control. The cost function for each next possible node (as the target node) consists of probability to closest *OCP*, P_{OCPi} , and probability to random point, P_{rand} .

In case of overlap between a selected node and obstacle in the environment, the selected node is discarded, and a new node is selected based on P_{OCPi} and P_{rand}. Setting the probabilities as P_{OCPi} =0.9 and P_{rand}=0.1, yield to the exploration behavior presented in Figure 20.

3.1.1 STP Planner Pseudo-Code

We present our STP planner pseudo code described in Table II, for spatial case generating trajectory T with search space method presented in the Section V.B. The search space is based on POCPi and Prand. We apply K steps from initial state xinit. The *f* function defines the dynamic model and kinematic constraints, $\dot{x} = f(x; u)$, where u is the input and OCPi are local target points between start to goal states.

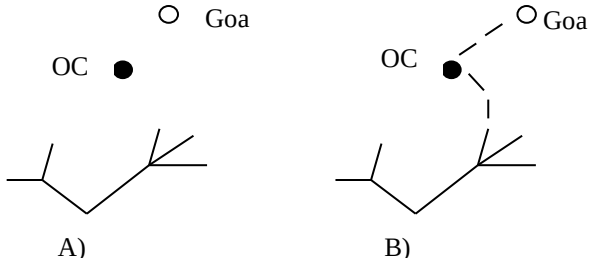


Figure 4. STP Search Method: (A) Start and Goal Points; (B) Explored Space to the Goal Through OCP

TABLE II. STP PLANNER PSEUDO CODE

STP Planner (x_{init}; x_{Goal} ;K; Δt; OCP)
T.init (x_{init});
$x_{rand} \leftarrow$ *random.state*();
$x_{near} \leftarrow$ *nearest.neighbor*(x_{rand}; *T*);
$u \leftarrow$ *select.input*(x_{rand}; x_{near});
$x_{new} \leftarrow$ *new.state.OCP* (OCP_1; *u*; Δ*t*; *f*);
While $x_{new} \neq x_{Goal}$ do
    $x_{rand} \leftarrow$ *random.state*();
    $x_{near} \leftarrow$ *nearest.neighbor*(x_{rand}; *T*);
    $u \leftarrow$ *select.input*(x_{rand}; x_{near});
    $x_{new} \leftarrow$ *new.state.OCP* (OCP_i; *u*; Δ*t*; *f*);
    T.add.vertex(x_{new});
    T.add.edge(x_{near}; x_{new}; *u*);
end
return *T;*

Function new.state.OCP (OCP_i;*u*; Δ*t*; *f*)
Set P_{OCPi} , *Set* P_{rand}
p←*uniform_rand[0..1]*
    if $0 < p < P_{OCPi}$
    return $x_{new} = f(OCP_i,u,\Delta t)$;
else
    if $P_{OCPi} < p < P_{rand} + P_{OCPi}$
then
    return RandomState();
    end.

C. Completeness

Motion-planning and search algorithms commonly describe 'complete planner' as an algorithm that always provides a path planning from start to goal in bounded time. For random sampling algorithms, 'probabilistic complete planner' is defined as: if a solution exists, the planner will eventually find it by using random sampling. In the same manner, the deterministic sampling method (for example, grid-based search) defines completeness as resolution completeness.

Sampling-based planners, such as the STP planner, do not explicitly construct search space and the space's boundaries, but exploit tests with preventing collision with obstacles and, in our case, taking spatial considerations into

account. Similarly, to other common RRT planners, which share similar properties with the STP planner, our planner can be classified as a probabilistic complete one.

IV. STP-IRL ALGORITHM

In most Reinforcement Learning (RL) systems, the state is basically agent's observation of the environment. At any given state the agent chooses its action according to a policy. Hence, a policy is a road map for the agent, which determines the action to take at each state. Once the agent takes an action, the environment returns the new state and the immediate reward. Then, the agent uses this information, together with the discount factor to update its internal understanding of the environment, which, in our case, is accomplished by updating a value function. Most methods are using the use well-known simple and efficient greedy exploration method maximizing Q-value.

In case of velocity planning space as part of spatial analysis planning, each possible action is a possible velocity in the next time step, that also represent a viewpoint. The Q-value function is based on greedy search velocity, with greedy local search method. Based on that, TD and SARSA methods for RL can be used, generating visible trajectory in 3D urban environment.

A. Markov Decision Processes (MDP)

The standard Reinforcement Learning set-up can be described as a MDP, consisting of:

- **A finite set of states** *S*, comprising all possible representations of the environment.
- **A finite set of actions** *A*, containing all possible actions available to the agent at any given time.
- **A reward function** $R = \psi(s_t, a_t, s_{t+1})$, determining the immediate reward of performing an action at from a state s_t, resulting in s_{t+1}.
- **A transition model** $T(s_t, a_t, s_{t+1}) = p(s_{t+1} | s_t, a_t)$, describing the probability of transition between states s_t and $s_{t+1}$when performing an action a_t.

B. Temporal Difference Learning

Temporal-difference learning (or TD) interpolates ideas from Dynamic Programming (DP) and Monte Carlo methods. TD algorithms are able to learn directly from raw experiences without any particular model of the environment.

Whether in Monte Carlo methods, an episode needs to reach completion to update a value function, Temporal-difference learning is able to learn (update) the value function within each experience (or step). The price paid for being able to regularly change the value function is the need to update estimations based on other learnt estimations (recalling DP ideas). Whereas in DP a model of the environment's dynamic is needed, both Monte Carlo and TD approaches are more suitable for uncertain and unpredictable tasks.

Since TD learns from every transition (state, reward, action, next state, next reward) there is no need to ignore/discount some episodes as in Monte Carlo algorithms.

C. STP Using Inverse Reinforcement Learning

In this section, we present Inverse Reinforcement Learning (IRL) approach based on the proposed Spatial RRT planning method. It considers that the value function *f* related to each point *x*. The Spatial RRT planner seeks to obtain the trajectory *T** that based on visibility motion primitives set by SVC Optimal Control Points (OCP) as part of the planned trajectory, which takes into account exact 3D visible volumes analysis clustering in urban environments, based on optimizing value function *f* along *T*.

The generated trajectories are then represented by a set of discrete configuration points $T = \{x_1, x_2, \cdots, x_N\}$. Without loss of generality, we can assume that the value function for each point can be expressed as a linear combination of a set of sub-value functions, that will be called features $c(x) = \sum c_j f_j(x)$. The cost of path *T* is then the sum of the cost for all points in the path. Particularly, in the RRT, the value is the sum of the sub-values of moving between pairs of states in the path:

$$c(\zeta) = \sum_{i=1}^{N-1} c(x_i, x_{i+1}) = \sum_{i=1}^{N-1} \frac{c(x_i) + c(x_{i+1})}{2} \|x_{i+1} - x_i\| = \omega^T \sum_{i=1}^{N-1} \frac{f(x_i) + f(x_{i+1})}{2} \|x_{i+1} - x_i\| = \omega^T f(\zeta) \quad (2)$$

Based on number of demonstration trajectories D, $D = \{\zeta_1, \zeta_2, \cdots, \zeta_D\}$, by using IRL, weights ω can be set for learning from demonstrations and setting similar planning behavior. As was shown by (Abbeel and Ng 2014; Kuderer et al. 2015), this similarity is achieved when the expected value of the features for the trajectories generated by the planner is the same as the expected value of the features for the given demonstrated trajectories:

$$\mathbb{E}(f(\zeta)) = \frac{1}{D} \sum_{i=1}^{D} f(\zeta_i) \quad (3)$$

Applying the Maximum Entropy Principle (Ziebart et al. 2008) to the IRL problem leads to the following form for the probability density for the trajectories returned by the demonstrator:

$$p(\zeta|\omega) = \frac{1}{Z(\omega)} e^{-\omega^T f(\zeta)} \quad (4)$$

where Z(ω) is a normalization function that does not depend on ζ. One way to determine ω is maximizing the (log-) likelihood of the demonstrated trajectories under the previous model:

$$L(D|\omega) = -D\log(Z(\omega)) \quad +\sum_{i=1}^{D} (-w^T f(\zeta_i)) \qquad (5)$$

The gradient of the previous log-likelihood with respect to ω is given by:

$$\nabla\mathcal{L} = \frac{\partial\mathcal{L}(\mathcal{D}|\omega)}{\partial\omega} = \mathbb{E}(f(\zeta)) - \frac{1}{D}\sum_{i=1}^{D} f(\zeta_i) \qquad (6)$$

As mentioned in (Kuderer et al. 2015), this gradient can be intuitively explained. If the value of one of the features for the trajectories returned by the planner are higher from the value in the demonstrated trajectories, the corresponding weight should be increased to increase the value of those trajectories.

The main problem with the computation of the previous gradient is that it requires to compute the expected value of the features E($f(\zeta)$) for the generative distribution (4).

We suggest setting large amount of D cased, setting the relative *w* values for our planner characters.

TABLE III. STP-IRL PLANNER PSEUDO CODE

STP - IRL Planner
Setting Trajectory S Examples D, D= T.init* (x_{init});
Calculate function features Weight, w
$f_D \leftarrow$ *AverageFeatureCount(D);*
w ← *random_init();*
Repeat
  for each T do*
    for rrt_repetitions do
      $\zeta_i \leftarrow$ *getRRTstarPath(T*,ω)*
      $f(\zeta_i) \leftarrow$ *calculeFeatureCounts*(ζ_i)
    end for
  f_{RRT} *(T*)*$\leftarrow\sum_{i=1}^{rrt\_repetitions} f(\zeta_i)$*/rrt_repetitions*
  end for
  $f_{RRT} \leftarrow (\sum_{i=1}^{S} f_{RRT})/s$
  $\nabla L \leftarrow f_{RRT} - f_D$
  w ←*UpdatedWeigths (*∇L)
Until convergence
Return w

V. PATROLLING PLANNING USING STP

In this section, we study the visible trajectories planning for patrolling application using heterogeneous multi agents in 3D urban environments. Our concept is based spatial clustering method using visibility analysis of the 3D visibility problem from a viewpoints in 3D urban environments, defined as locations. We consider two kinds of agents, with different kinematic and perception capabilities. Using simplified version of Traveling Salesman Problem (TSP), we formulate the problem as patrolling strategy one, with upper bound optimal performances. We present combination of relative deadline UniPartition approaches based on visibility clusters. These key features allow new planning optimal patrolling strategy for heterogeneous agents in urban environment. We demonstrate our patrolling strategy method in simulations using Autonomous Navigation and Virtual Environment Laboratory (ANVEL) test bed environment.

VI. PROBLEM FORMULATION

The definition of the problem commonly set according to a predefined strategy planning. Patrolling strategy deals with different situations and hence yields different formulation. We demonstrate these differences using the following division:

1. Force Planning - with a given set of locations that should be protected, we have to determine what is the minimal number of agents (K) with patrolling strategy which meets all constraints. That becomes the common case, when we do not have any available agents compatible to the patrolling mission and we need to decide how many agents should be used in order to meet mission constraints.
2. Force Division - The locations are situated in different areas in urban environment. We need to decide how to allocate the agents to all the areas in such way that it will be possible to find a patrolling strategy that meets all constraints in all the areas.

In our case, we deal with force division problem combining relative deadline and visibility clustering. Given a set of N locations and K different types of agents (which available for patrol in a given moment), we are focusing on finding patrolling strategy, where each route for an agent passes through a number of locations. Patrolling strategy aims to minimize cost function which based on 3D visible volumes and meets the relative deadline constraints.

A. Locations

We set on our urban environment number of assets that should be protected, named as our Locations. Each location is characterized by:

1. Coordinates - its actual location in the environment generated from Visibility Clustering based on 3D visible volumes analysis.
2. Required Protection - different locations have different characteristics and therefore may differ in their protection needs. As will be discussed later, we refer to the type of agent required for protection.
3. Relative Deadline - which is the maximum time that may pass between consecutive visits for optimal

patrolling (for example, reduce the probability for being attacked beneath a certain threshold).

B. Agents

Our Agents are modeled as Unmanned Ground Vehicles (UGVs). We will generalized and differentiate between two types of these kind of agents:

1. Large - Agent with long range patrolling capabilities, sensing and target engaging abilities.
2. Small – Agent with limited ranges in the aspects described above.

Each of the agent's type has different average speed, dynamic and kinemtic constraints with hourly operational cost of its own and perception capabilities related to visibility analysis. There are many other factors to be considered but it is also beyond the scope of this paper.

VII. PATROLLING STRATEGY

We define collection of routes for all agents (one route per agent) which cover all the locations, by one agent or another, as Patrol Strategy. The total patrol cost defined as sum of all the routes cost, where route cost includes 3D visible volumes aspect integrated into visibility clusters.

A. Main Assumptions

We have assumed the following assumptions:

1. Agent Type - Each location requires different type of protection, and each type of agent has unique and different capabilities. Therefore, we will assume that each location can be protected by only one type of agent: large or small one. These assumptions split our problem into two independent problems, each with its relevant locations and relative deadlines.
2. Disjoint routs for each agent - In order to simplify our problem, we assume that each agent is allocated to patrol on a specific subset of locations. The subsets are disjoint sets which create disjoin routes for each agent. The union of those subsets is the initial set of locations.
3. Single visit to locations - Another simplification of the problem is the assumption that along the route each location can be visited only once, and the problem can be described as TSP private case. Due to this assumption, when we have locations with a relatively short deadlines, there might be cases where no solution will be found. This will be addressed in the review of the solution technique.
4. Patrolling along a cyclic path - We assume that each patrol route is a cyclic path. The definition of cyclic path is a path which start and end at the same location.
5. Traveling time between locations - Shortest traveling time - The given traveling times are the shortest possible times between given two locations, based on agent's dynamic model.

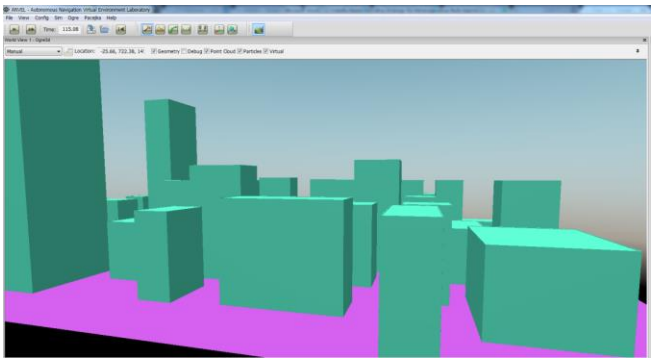
B. Program Inputs

The inputs of the program are as follows:

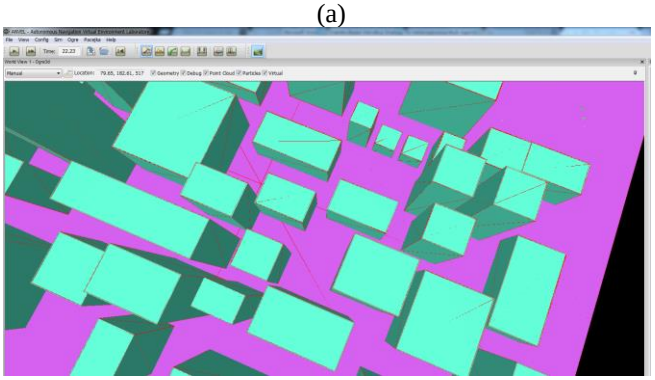
- N- number of locations
- K- number of agents
- Size of patrolling area
- Agent's average speed and physical dimensions
- C- The cost per one operational hour
- F- The cost per one hour of deviation from the deadlines (the fine cost)

We assumed 10 locations in total (N=10) and 3 agents in total (K=3), C=100$, F=1000$ and average speed of the agent is 65 km/h. We randomly generated the relative deadline of each location (9h in average) and also the coordinates of the N locations in the patrolling area (500 km^2).

The problem has the following dimensions: ten locations for small agents, six locations for large agents, and three small agents patrol. Our testbed urban environment can be seen in Figure 5. We set two large agents on patrol and one small agent in a given time. Agents in simulation environment can be seen in Figure 6. Perception capabilities of each agent can be seen in Figure 7 and Figure 8. Locations are changed according to the initial position of the agents as each simulation. The locations are set as the outcome of our visibility clustering analysis as can be seen in Figure 9 and Figure 11.



(a)



(b)

Figure 5. Our test bed urban environment model in ANVEL, (a) Sideview; (b) Topview.

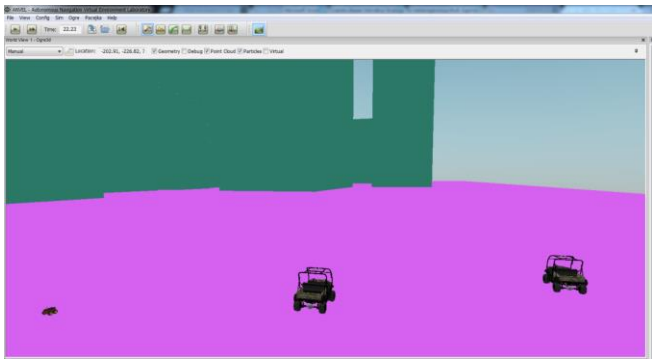


Figure 6. Three agents in our simulation, two large agents on the right side and a small agent to the left.

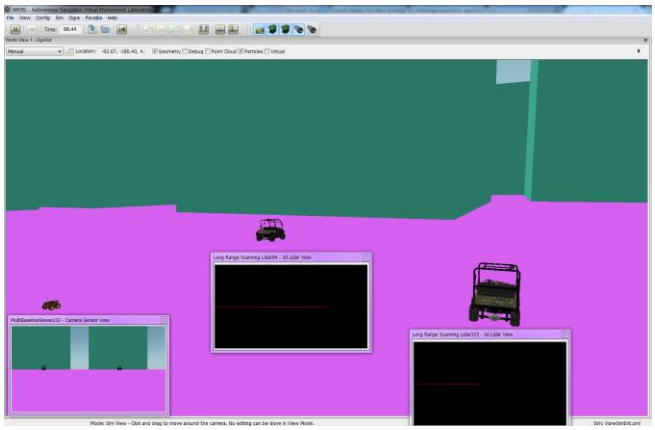


Figure 7. Perception sensor; LIDAR sensor for large agents and camera sensor for a small agent.

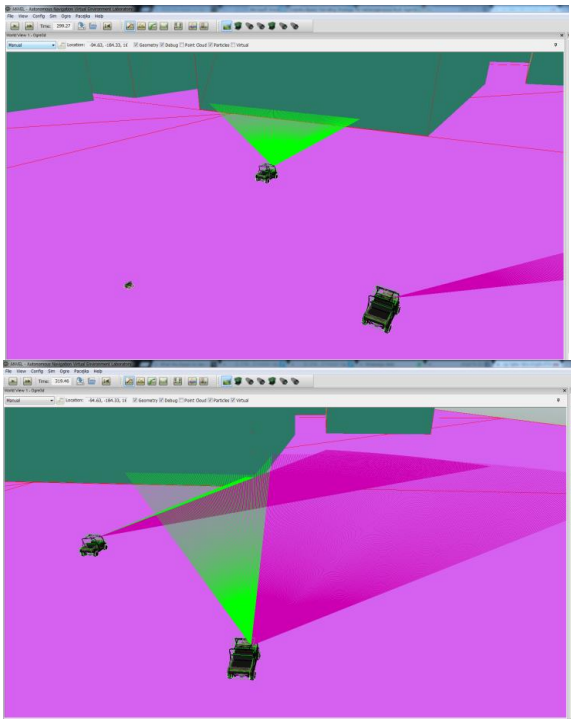


Figure 8. LIDAR perception capabilities during time

C. Simulations Results

We will preset several examples of the program results, based on spatial clustering and patrolling strategy.

In the first case, coordinates of the 10 locations are presented in Figure 9.

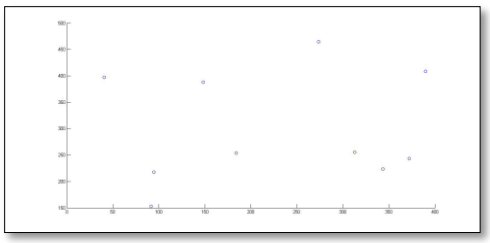


Figure 9. Location Coordinates based on Visibility Clustering

The patrolling graph can be seen in Figure 10.

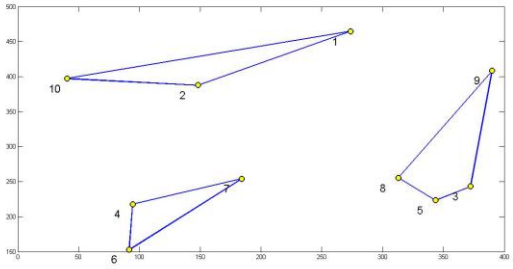


Figure 10. Patrolling Strategy Trajectories According to Problem Constraints

Patrol #1 [10, 1, 2, 10], Patrol #2 is [6, 4, 7, 6] and Patrol #3 [8, 5, 3, 9, 8]. Total time of the solution is (h) 48.885, and the total deviation from deadlines is (h) 0. Total cost of the solution is 4888.5. The total time of the solution is the sum of the total traveling time of all three patrols. For example, the total traveling time of patrol #3 is the time to arrive and return to location 8 plus the traveling time between the locations in the cluster, meaning the traveling time from location 8 to 5, 5 to 3, 3 to 9 and 9 to 8.

Our second simulation demonstrate patrol strategies that are non-feasible. Coordinates of the 10 locations are presented in Figure 11.

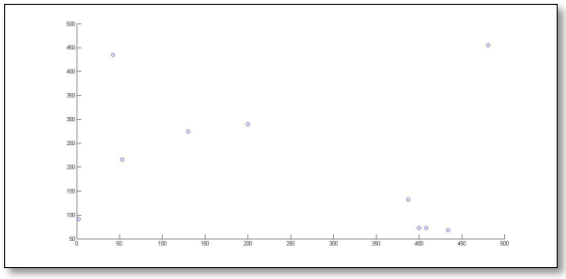


Figure 11. Location Coordinates based on Visibility Clustering in Second Simulation

The patrolling graph can be seen in Figure 12.

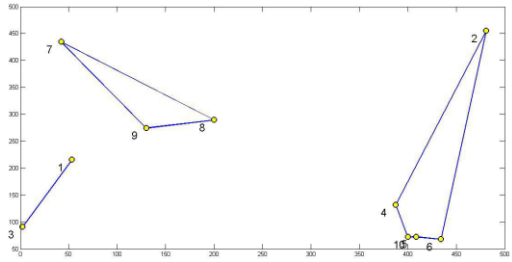


Figure 12. Patrolling Strategy Trajectories According to Problem Constraints

The program prints the following results: Patrol #1 [3, 1, 3], Patrol #2 [10, 4, 2, 6, 5, 10], Patrol #3 [9, 7, 8, 9], as can be seen in Figure 12. Total time of the solution is (h) 48.65, total deviation from deadlines is (h) 4.53. The total deviation from deadlines is the sum of the deviations of the patrol strategies that are non-feasible (it is not necessary that all the 3 patrol strategies have a deviation).

Other interesting example of a case in which one of the patrol strategy containing only one location. The reason for that is the coordinates of this location which are far away from all the other locations.

In the first case, coordinates of the 10 locations are presented in Figure 13.

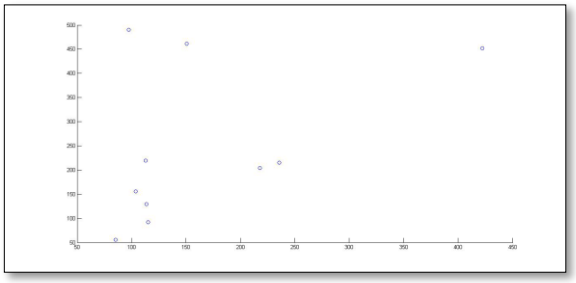


Figure 13. Location Coordinates based on Visibility Clustering in Third Simulation

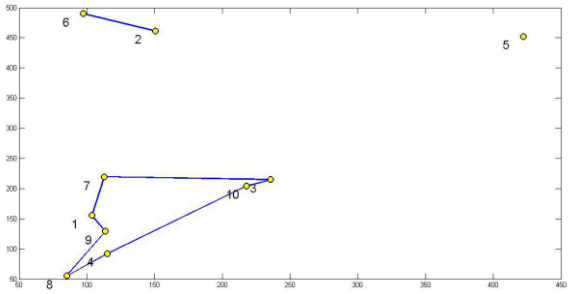


Figure 14. Patrolling Strategy Trajectories According to Problem Constraints

The program prints the following results: Patrol #1 [8, 9, 1, 7, 3, 10, 4, 8], Patrol #2 [2, 6, 2], Patrol #3 [5], as can be seen in Figure 14. Total time of the solution is (h) 46.88, total deviation from deadlines is (h) 0.

D. Discussion

Our research tackled simplified version (given our assumptions) to a problem as described in the introduction. In this section we would like to discuss other complicated cases which are received by eliminating one or more of our assumptions and cases which are received by different problem definition. We suggest a general solution approach for those cases and summarize with recommendation for possible research extensions.

1. **Finding a patrol strategy when repeated visits are allowed** - as we described in the second step of our solution technique (detailed in section (4)), the patrol strategy in each cluster will be found by using TSP technique which does not allow repeated visits apart from the first location in the path. Naturally, there might be a path with repeated visits which is a feasible solution (all the relative deadlines are met), but the TSP technique will "miss" this solution. Using the TSP technique is one of the alleviations that we have made, and we are aware to the fact that we might miss feasible solutions. In order to deal with the case in which there are no feasible solution at all, we added the fine calculation, but here we would like to shortly present a heuristic approach for finding a patrol strategy with repeated visits.

This approach is based on general idea as follows: For each cluster, the patrol strategy receives a feasible solution. Given a starting location (which is also the end location), the method uses the following algorithms:

- Forward Checking - given the last location in the patrol forming strategy this algorithm finds all the possible locations for that will satisfy the second and third constraints. All those locations are added to a location vector.

- Recursive call - given the last location in the patrol forming strategy and the index of the following step, it checks if the patrol have not ended with all constraints satisfied. If that is not the case, it calls Forward checking and run itself recursively on every location in and the index until a patrol strategy is achieved. Meaning, we have a route which is a closed path and every location was visited at least once. When the algorithm backtracks till a different location can be chosen and continue that route.

2. **Non disjoint routes** - one of the alleviations that we made is to assume that the route of each agent is disjoint from all other routes. The reason for that was to avoid collisions between the agents and to simplify our problem. Obviously, in reality, the routes does not have to be disjoint. When eliminating this assumption the complexity of the problem rises due to the fact that there are a great deal more cluster partition options. We suggest a heuristic approach to try and find a solution to the problem. First preform the

same partition of the location into clusters (or similar partition method based on geographical proximity) and then check if there is a solution with disjoint routes. If there is no such solution, find the locations that for them the relative deadline constraint is not met. Then, try and add each of those locations to another cluster based on a proximity criteria, such as the shortest average distance to all locations in the cluster and so forth. Now these locations are visited by more then one agent and therefore the time between consecutive visits is reduced to a point that may yield a feasible solution.

3. **Joint routes for a number of agents while the goal is to minimize the maximal idle time of all locations** - This is the approach mentioned in the introduction (see section (1)). Well known heuristic algorithms solve the following problem:

- Create an Outer-planner Graph from all given locations. This is done by connecting all the locations with arcs that does not cross one another. Allocate all agents to this graph and calculate the maximal idle time.

- By removing two arcs at a time from the graph we are creating two separated graphs. Remove all possible pairs of arcs and for each graph pair created find the agent allocation (meaning the number of agents allocated to each graph) that minimizes the maximal idle of both graphs.

- Find the division that yields the minimal maximal idle time and for each graph preform the same process described in the previous stage.

- Continue until the maximal idle time can not be reduced or when the only possible allocation is single agent to a graph.

4. **Locations that require protection from both type of agents** - in our solution, we assumed that each agent can be protected by only one type of agent. The reason for that was to separate the initial problem into two disjoin problem. That way, we can solve each problem separately. Obviously, in reality, there exist locations that require protection by both types of agents (we will refer to them as hybrid locations). While there are different relative deadlines for each type of agent then we can continue referring to the problem as two separated problems. The complication begins when the relative deadline is unified meaning the maximal time that may pass between consecutive visits of any type of agent. A possible way of approaching this case without unifying the problems (which may yield a large dimension problem that is harder to solve) is as follows:

- Solve the separated problems.

- If no feasible solution exists for one of the problems, and that is due only to a problem in meeting one or more hybrid locations relative deadline constraints. Continue to the next stage

- Combine the solution with the total minimal deviation from the relative deadlines of the hybrid locations with feasible solutions (or such with minimal deviation as well) for the other type of agents and update the time between consecutive visits to hybrid locations accordingly. Check if now the relative deadlines are met.

- If no feasible solution can be found that way, the fine method we introduced in this paper may be used.

VIII. CONCLUSIONS

In this paper, we have presented a unique planner concept, STP, generating trajectory in 3D urban environments based on UGV model. The planner takes into account obstacle avoidance capabilities and passes through optimal control points calculated from spatial analysis. The spatial analysis defines the number of clusters in a dataset based on an analytic visibility analysis, named SVC.

Based SVC and STP analysis, we presented an Inverse Reinforcement Learning (IRL) approach based on the proposed STP planning method, learning the value function of the planner from demonstrated trajectories.

We also presented visible trajectories planning for patrolling application using heterogeneous multi agents in 3D urban environments. Using ANVEL simulation environment, we demonstrated spatial clustering method using visibility analysis of the 3D visibility problem from a viewpoints in 3D urban environments. As part of our simulations we modeled two kinds of agents, with different kinematic and perception capabilities. Patrolling strategy formulated as Traveling Salesman Problem (TSP), with relative deadline UniPartition approaches based on visibility clusters.

We showed implementation of differences cases, using large and small agents in urban environment scenarios where sometime trajectory can no be found. Some other problem formulation discussed with suggested solution for further research, such as: finding a patrol strategy when repeated visits are allowed; non disjoint routes as part of our patrol graph; joint routes for a number of agents while the goal is to minimize the maximal idle time of all locations and other cases of locations that require protection from both type of agents.

Future research will also include performances and algorithm complexity analysis for STP and SVC methods.

IX. REFERENCES

O. Gal and Y. Doytsher, (2014) "Spatial Visibility Clustering Analysis In Urban Environments Based on Pedestrians' Mobility Datasets," The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services, pp. 38-44.

J. Bellingham, A. Richards, and J. How, (2002) "Receding Horizon Control of Autonomous Aerial Vehicles," in Proceedings ofthe IEEE American Control Conference, Anchorage, AK, pp. 3741–3746.

A. Borgers and H. Timmermans, (1996) "A model of pedestrian route choice and demand for retail facilities within inner-city shopping areas," Geographical Analysis, vol. 18, No. 2, pp. 115-128.

S. A. Bortoff, (2000) "Path planning for UAVs," In Proc. of the American Control Conference, Chicago, IL, pp. 364–368.

O. Brock and O. Khatib, (2000) "Real time replanning in high-dimensional configuration spaces using sets of homotopic paths," In Proc. of the IEEE International Conference on Robotics and Automation, San Francisco, CA, pp. 550-555.

R. B. Calinski and J. Harabasz, (1974) “A Dendrite Method for Cluster Analysis,” Communications in Statistics, vol. 3, pp. 1–27.

B. J. Capozzi and J. Vagners, (2001) “Navigating Annoying Environments Through Evolution,” Proceedings of the 40th IEEE Conference on Decision and Control, University of Washington, Orlando, FL.

H. Chitsaz and S. M. LaValle, (2007) "Time-optimal paths for a Dubins airplane," in Proc. IEEE Conf. Decision. and Control., USA, pp. 2379–2384.

B. Donald, P. Xavier, J. Canny, and J. Reif, (1993) “Kinodynamic Motion Planning,” Journal of the Association for Computing Machinery, pp. 1048–1066.

Y. Doytsher and B. Shmutter, (1994) "Digital Elevation Model of Dead Ground," Symposium on Mapping and Geographic Information Systems (Commission IV of the International Society for Photogrammetry and Remote Sensing), Athens, Georgia, USA.

F. Durand, (1999) "3D Visibility: Analytical Study and Applications," PhD thesis, Universite Joseph Fourier, Grenoble, France.

M. Erdmann and T. Lozano-Perez, (1987) "On multiple moving objects," Algorithmica, Vol. 2, pp. 477–521.

V. Estivill-Castro and I. Lee, (2000) "AMOEBA: Hierarchical Clustering Based on Spatial Proximity Using Delaunay Diagram," In Proceedings of the 9th International Symposium on Spatial Data Handling, Beijing, China.

P. Fiorini and Z. Shiller, (1998) "Motion planning in dynamic environments using velocity obstacles," Int. J. Robot. Res. vol. 17, pp. 760–772.

W. Fox, D. Burgard, and S. Thrun, (1997) "The dynamic window approach to collision avoidance," IEEE Robotics and Automation Magazine, vol. 4, pp. 23–33.

T. Fraichard, (1999) "Trajectory planning in a dynamic workspace: A ’state-time space’ approach," Advanced Robotics, vol. 13, pp. 75–94.

E. Frazzoli, M.A. Daleh, and E. Feron, (2002), "Real time motion planning for agile autonomous vehicles," AIAA Journal of Guidance Control and Dynamics, vol. 25, pp. 116–129.

O. Gal and Y. Doytsher, (2012) "Fast and Accurate Visibility Computation in a 3D Urban Environment," in Proc. of the Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services, Valencia, Spain, pp. 105-110.

P. Arabie and L. J. Hubert, (1996) "An Overview of Combinatorial Data Analysis," in Arabie, P., Hubert, L.J., and Soete, G.D. (Eds.) Clustering and Classification, pp. 5-63.

O. Gal and Y. Doytsher, "Fast Visibility Analysis in 3D Procedural Modeling Environments," in Proc. of the, 3rd International Conference on Computing for Geospatial Research and Applications, Washington DC, USA, 2012.

O. Gal and Y. Doytsher, (2013) "Fast Visibility in 3D Mass Modeling Environments and Approximated Visibility Analysis Concept Using Point Clouds Data," Int. Journal of Advanced Computer Science, IJASci, vol. 3, no. 4, April 2013, ISSN 2251-6379.

O. Gal and Y. Doytsher, (2013) "Fast and Efficient Visible Trajectories Planning for Dubins UAV model in 3D Built-up Environments," Robotica, FirstView, Article pp. 1-21 Cambridge University Press 2013.

A. Gordon, (1999) Classification (2nd ed.), London: Chapman and Hall/CRC Press.

S. Guha, R. Rastogi, and K. Shim, (1998) "CURE: An efficient clustering algorithm for large databases," In Proceedings of the ACM SIGMOD Conference, Seattle, WA, pp. 73-84.

M. Haklay, D. O’Sullivan, and M.T. Goodwin, (2001) "So go down town: simulating pedestrian movement in town centres," Environment and Planning B: Planning & Design, vol. 28, no. 3, pp. 343-359.

D. Harel and Y. Koren, (2001) "Clustering spatial data using random walks," In Proceedings of the 7th ACM SIGKDD, San Francisco, CA, pp. 281-286.

J. Hartigan, (1975) "Clustering Algorithms". John Wiley & Sons, New York, NY.

J. Hartigan and M. Wong, (1979) "Algorithm AS136: A k-means clustering algorithm," Applied Statistics, vol. 28, pp. 100-108.

S. P. Hoogendoorn and P. H. L. Bovy, (2001) "Microscopic pedestrian way finding and dynamics modelling," In Schreckenberg, M., Sharma, S.D. (eds.) Pedestrian and Evacuation Dynamics. Springer Verlag: Berlin, pp. 123-154.

D. Hsu, R. Kindel, J-C. Latombe, and S. Rock, (2000) "Randomized kinodynamic motion planning with moving obstacles," Algorithmics and Computational Robotics, vol. 4, pp. 247–264.

B. Jiang, (1999) "SimPed: Simulating pedestrian flows in a virtual urban environment," Journal of Geographic Information and Decision Analysis, vol. 3, no. 1, pp. 21-30.

S. Karaman and E. Frazzoli, (2011) “Sampling-based algorithms for optimal motion planning,” Int. J. Robot. Res., vol. 30, no. 7, pp. 846–894.

S. Karaman, M. Walter, A. Perez, E. Frazzoli, and S. Teller, (2011) “Anytime motion planning using the RRT*,” in Proc. IEEE Int. Conf. Robot. Autom., Shanghai, pp. 1478–1483, May.

L. Kaufman and P. Rousseeuw, (1990) "Finding Groups in Data: An Introduction to Cluster Analysis," John Wiley and Sons, New York, NY.

N.Y. Ko and R. Simmons, (1998) "The lane-curvature method for local obstacle avoidance," In International Conference on Intelligence Robots and Systems.

W. J. Krzanowski and Y. T. Lai, (1985) “A Criterion for Determining the Number of Groups in a Data Set Using Sum of Squares Clustering,” Biometrics, vol. 44, pp. 23–34.

M.P. Kwan, (2000) "Analysis of human spatial behavior in a GIS environment: recent developments and future prospects," Journal of Geographical System, no. 2, pp. 85-90, 2000.

J. C, Latombe, (1990) "Robot Motion Planning,", Kluwer Academic Press.

S. M. LaValle, (1998) "Rapidly-exploring random trees: A new tool for path planning," TR 98-11, Computer Science Dept., Iowa State University.

S. M. LaValle, (2006) "Planning Algorithms," Cambridge, U.K.: Cambridge Univ. Press.

S. M. LaValle and J. Kuffner. (1999) "Randomized kinodynamic planning," In Proc. IEEE Int. Conf. on Robotics and Automation, Detroit, MI, pp. 473–479.

L.R. Lewis, (2006) "Rapid Motion Planning and Autonomous Obstacle Avoidance for Unmanned Vehicles," Master's Thesis, Naval Postgraduate School, Monterey, CA, December.

C. W. Lum, R. T. Rysdyk, and A. Pongpunwattana, (2006) “Occupancy Based Map Searching Using Heterogeneous Teams of Autonomous Vehicles,” Proceedings of the 2006 Guidance, Navigation, and Control Conference, Autonomous Flight Systems Laboratory, Keystone, CO, August.

G.W. Milligan and M. C. Cooper, (1985) “An Examination of Procedures for Determining the Number of Clusters in a Data set,” Psychometrika, vol. 50, pp. 159–179.

J. Minguez and L. Montano, (2000) "Nearest diagram navigation. a new realtime collision avoidance approach," In International Conference on Intelligence Robots and Systems.

J. Minguez, N. Montano, L. Simeon, and R. Alami, (2002) "Global nearest diagram navigation," In Proc. of the IEEE International Conference on Robotics and Automation.

B. Moulin, W. Chaker, and J. Perron, (2003) "MAGS project: Multi-Agent GeoSimulation and Crowd Simulation," Kuhn, W., Worboys, M.F. and Timpf, S. (Eds.): LNCS 2825, pp. 151–168.

K. J. Obermeyer, (2009) “Path Planning for a UAV Performing Reconnaissance of Static Ground Targets in Terrain,” in Proceedings of the AIAA Guidance, Navigation, and Control Conference, Chicago.

S. Okazaki and S. Matsushita, (1993) "A study of simulation model for pedestrian movement with evacuation and queuing," Proceedings of the International Conference on Engineering for Crowd Safety, London, UK, pp. 17-18.

A.Pongpunwattana and R.T. Rysdyk, (2004) “Real-Time Planning for Multiple Autonomous Vehicles in Dynamic Uncertain Environments,” AIAA Journal of Aerospace Computing, Information, and Communication, pp. 580–604.

H. Plantinga and R. Dyer, (1990) "Visibility, Occlusion, and Aspect Graph," The International Journal of Computer Vision, vol. 5, pp. 137-160.

J. Sasiadek and I. Duleba, (2000) "3d local trajectory planner for uav," Journal of Intelligent and Robotic Systems, vol. 29, pp. 191–210.

V. Shaferman and T. Shima, (2008) “Co-evolution genetic algorithm for UAV distributed tracking in urban environments,” in ASME Conference on Engineering Systems Design and Analysis.

T. Schelhorn, D. Sullivan, and M. Haklay, (1999) "STREETS: An agent-based pedestrian model,".

C. Stachniss and W. Burgard, (2002) "An integrated approach to goal directed obstacles avoidance under dynamic constrains for dynamic environment," In International Conference on Intelligence Robots and Systems.

R. Tibshirani, G. Walther, and T. Hastie, (2001) “Estimating the Number of Clusters in a Dataset via the Gap Statistic,” Journal of the Royal Statistical Society, Ser. B, vol. 32, pp. 411–423.

L. Ulrich and J. Borenstien, "Vfh+: Reliable obstacle avoidance for fast mobile robots," In Proc. of the IEEE International Conference on Robotics and Automation, 1998.

Abbeel, P., Ng, A.Y., (2004) "Apprenticeship learning via inverse reinforcement learning" In: Proceedings of the twenty-first international conference on Machine learning, ICML ’04, ACM, New York, NY, USA.

Kuderer, M., Gulati, S., Burgard, W, (2015) "Learning driving styles for autonomous vehicles from demonstration", In: Proceedings of the IEEE International Conference on Robotics & Automation (ICRA), Seattle, USA. vol. 134.

Ziebart, B., Maas, A., Bagnell, J., Dey, A., (2008) "Maximum entropy inverse reinforcement learning", In: Proc. of the National Conference on Artificial Intelligence (AAAI).

Signal Processing in Vibration Analysis with Application in Predictive Maintenance of Rotating Machines

Theodor D. Popescu

National Institute for Research
and Development in Informatics
8-10 Averescu Avenue
011455 Bucharest, Romania
Email: Theodor.Popescu@ici.ro

Dorel Aiordăchioaie and Anisia-Culea Florescu

"Dunărea de Jos" University of Galaţi
47 Domneasca Street
800008 Galaţi, Romania
Email: Dorel.Aiordachioaie@ugal.ro
Email: Anisia.Florescu@ugal.ro

***Abstract*—A general approach for change detection in vibration signals with application in predictive maintenance of rotating machines represents the object of the paper. After an overview of the maintenance approach, the condition monitoring in predictive maintenance is presented. Also, some vibration analysis techniques, making use of change detection, independent component analysis, time-frequency analysis and energy distribution, with application in predictive maintenance of rotating machinery, are discussed. They can be combined in a unified approach, offering new possibilities for more robust detection of changes in vibration signals, and assuring proactive actions in predictive maintenance. Finally, some experimental results for detection of faults in rolling element bearings and in a rotating machine operating, an industrial pump, are presented.**

***Index Terms*—Predictive maintenance; Change detection; Independent component analysis; Time-frequency analysis; Energy distribution; Rolling element bearings; Industrial pump.**

I. Introduction

Vibration analysis is the one of the most effective tool used to check the health of plant machinery and diagnose the causes. The health of machine is checked by routine or continuous vibration monitoring with sophisticated instruments, giving an early indication of a possible failure and offering countermeasures to avoid a possible catastrophic event. The paper presents some vibration analysis techniques, and different combination of these, in order to offer a framework for predictive maintenance of rotating machines and represents an extended and enhanced version of [1].

Vibration monitoring problem consists of machines condition and the change rate of its behavior. It can be ascertained by selecting of a suitable parameter for deterioration measuring and recording its value for further analysis. This activity is known as condition monitoring. The great parts of the defects encountered in the rotating machinery give rise to a distinct vibration pattern, or "vibration signature". Vibration monitoring has the ability to record and identify vibration "signatures" for monitoring rotating machinery. Vibration analysis is applied by using transducers to measure acceleration, velocity or displacement, depending of the frequencies making the object of the analysis. Careful scrutiny and deep study of vibration "signature" eliminate different fault possibilities and concludes to single fault. A logical and systematic approach has proved successful in diagnosing the basic causes. This applies to small, medium, large, direct coupled machines, motors, pumps, generators, turbo-machinery fans and compressors. Some machines are directly coupled to motor or some through gear boxes.

Sometimes, the vibration monitoring makes use of different change detection (CD) techniques. From statistical point of view, these techniques identify changes in the probability distribution of a stochastic process. The problem involves both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifies the times of such changes produced. These techniques can be classified either as time, frequency or time-frequency domain based algorithms. They are based on distance measures, artificial intelligence, fuzzy logic, statistical differences, etc., applied on the original signals or on the preprocessed signals, in order to amplify the changes in their dynamics. Presently, the CD problem represents a key point, when preventive maintenance is replaced by predictive maintenance.

Some features, among the amplitude levels in the time domain, are easily extracted and classified, but they are affected by noise. Energy distribution in the time-frequency domain [2], involving more operations, can lead to more robust change detection in vibrating signal dynamics. Also, parametric signal processing algorithms can be used for change detection if there is an accurate model of the signal, in a selected representation space. However, the approach, based on modeling techniques, has limitations as well.

The time-frequency analysis (TFA) [3], in comparison with the time-domain analysis, usually provides a simpler interpretation and comprehension of nonstationary signals, with large application in vibration monitoring. The idea is to analyze the behavior of the energy distribution (ED), i.e., the distribution of energy at certain instant or certain frequency band or more generally [2], in some particular time and frequency region. The results can represent a starting point in solving CD problems. So, new analysis facilities in CD problem solving,

are offered by the usage of the entropy based measures, such as Kullback-Leibler distance, Rényi distance, and Jensen difference, adapted to the time-frequency plane [4].

The paper is organized as follows. Section II refers to maintenance approach, while Section III presents the condition monitoring problem in predictive maintenance. Section IV has as subject change detection in vibration monitoring and is followed, in Section V, by a general view on the main signal processing techniques involved in vibration monitoring, with application in predictive maintenance of rotating machines. Section VI discusses different approaches in change detection of vibration signals based of signal processing techniques presented in Section V. Finally, in Section VII are discussed two case studies having as object fault detection in rolling element bearings (REB), and in a rotating machine, an industrial pump.

II. Maintenance Approach

Usually, the maintenance is performed as *preventive maintenance*, at fixed time intervals, or as *reactive maintenance*, when an actually fault produced. In the last case, it is necessary to perform immediately maintenance actions, while in the *predictive maintenance*, after a warning of a fault producing, the problem solving is carried out when necessary, so to avoid disruption of machine operating. A comparison of different maintenance types, with disadvantages and advantages is given in [5]. We present in the following some aspects concerning these approaches, to be taken into account, mainly in predictive maintenance of rolling element bearings.

A. Reactive Maintenance

This approach refers to machine running till a fault produced and involves fixing problems only the fault occurs. It represents the simplest and cheapest approach in terms of maintenance costs; often it implies additional costs, usually due to unplanned downtime. It can be seen as an easy solution to many maintenance strategies.

In rotating machines, rolling element bearings represent the most critical components, both in terms of initial selection and as well as in how they are maintained. Monitoring the condition of rolling bearings are essential and vibration based monitoring is frequently used to detect an early fault.

B. Preventive Maintenance

The preventive maintenance implies the scheduling of regular machine shutdowns, even if they are non required; this will increase the maintenance costs as some machine components are replaced, when this is not necessarily required. Some risks could appear due to replacing a defective machine part, incorrectly installing or reassembling parts. A frequently result of preventive maintenance consist of the fact that the maintenance is performed when there is nothing wrong in machine operating. Significant costs saving can be obtained by predictive maintenance.

C. Predictive Maintenance

The predictive maintenance refers to the process of monitoring the machine condition as it operates in order to predict which components are likely to fail and when. So, the maintenance can be planned and there is the possibility to change only those components that produce failure signs in its operating. The predictive maintenance principle consist of take measurements, to be used for prediction of the machine components behavior, susceptible of failure, and when these will be produced. Usually, these measurements include machine vibration, and machine operating parameters: flow, temperature, pressure, etc.

The continuous monitoring detects, in advance, the onset of component problems, so the maintenance is performed when needed. By this approach, unplanned downtime is reduced, as well as the risk of catastrophic failure. This will increase the efficiency and reducing of the costs. By predictive maintenance strategy, applied in rolling bearings, the costs can be avoid, giving in advance, a warning of a possible failure, enabling remedial action in advance.

III. Condition Monitoring

Condition monitoring consists of machine monitoring for early sign of failure so that the maintenance activity can be better planned, with reduced down time and costs.

The monitoring of vibration, temperature, voltage or current and oil analysis is frequently the most used. Vibration is the most widely used for its ability to detect and diagnose failure problems, but it offers also a prognosis on the useful life and possible failure mode of the machine. The prognosis is much more difficult to be performed and usually relies on continue monitoring of the fault to estimate the time when the machine will become unusable, taking into account the known experience in similar cases.

Vibration monitoring can be considered the most widely used predictive maintenance technique, and can be applied to a wide area of rotating machines. Machine vibration comes from many sources such as bearings, gears, unbalance etc., each sources having its own characteristic frequencies, manifesting as a discrete frequency, or as a sum and/or difference frequency. It can result complex vibration signals which put problems in vibration analysis, but some techniques, with a high sensitivity to faults, can reduce the complexity of the analysis. Bearing defects can affect higher frequencies, offering a basis for detecting incipient failure.

Usually, the detection uses the basic form of vibration measurement, where the vibration level is measured on a broadband basis (10-1000 Hz or 10-10000 Hz). The spikiness of the vibration signal, in machines with little vibration other than in the case of the bearings, is highlighted by the Crest Factor, indicating an incipient defect; also a great value of the energy, given by RMS level, indicates a severe defect.

Only this type of measurement offers limited information, but it can be useful for trend evaluation; increasing vibration

level highlights the machine condition deterioration. Also, a comparison of the measurement level with some vibration criteria from literature proves to be useful in practice.

Generally, rolling bearings produce very little vibration in faults absence, and present specific frequencies when a fault produced. At the beginning of a fault, for a single defect, the vibration signals present a narrow band frequency spectrum. As the malfunction increases, it can be noted an increase in the characteristic defect frequencies and sidebands, with a drop in these amplitudes, broadband noise increasing and considerable vibration at shaft rotational frequency [5]. At very low machine speed, low energy signals are generated by the bearings, difficult to be detected. Also, bearings located within a gearbox are difficult to monitor, because of the high energy at the gear, which can mask the bearing defect frequencies.

IV. Change Detection in Vibration Monitoring

The CD problem is frequently present for continuous monitoring of systems like machinery, structure, process, equipment or plant, using data provided by the sensors. So, it is possible to anticipate the abnormal functioning or these systems, before it is produced and to reduce the maintenance costs. The normal behavior of a system can be described by a parametric model, without using artificial excitation, reducing the speed of the equipment or temporary stop. If such early detections are possible, large changes of the system can be prevented, and the effects of defects, mechanical fatigue, etc. can be quickly anticipate, raising the availability of the system.

The applications in this field make use of theories based on statistics, providing theoretical instruments to solve the early detection problem. Many industrial processes are based on known physical principles, with available analytical models, and for very complicated or unknown models, semi-physical or black-box models can be used. Vibrations analysis and surveillance of machinery or industrial equipments represent important cases of detection and diagnosis problems.

The CD problem refers to detection of the change (the alarm) and evaluation of the change (estimation), providing information, in some case, for diagnosis (source isolation). The performance criterion of a change detection algorithm consists of its ability to correctly detect the changes, with minimum delay and minimum probability of false decisions. So, it must respond to the small changes (sensitivity to changes), and does not be affected by the disturbances, noise or modeling errors (robustness of the algorithm). The sensitivity and robustness properties are usually in conflict, a good change detection algorithm must perform a compromise between the two aspects.

Two basic approaches in CD are reported as based on quantitative models (using analytical redundancy) and qualitative models, which can be conveniently combined to improve the robustness of the generation of the quantitative residuals. In the case of analytical exact models absence, learning models, such as fuzzy and neural models, can be used. More, the neural networks can be used for classification of the residuals, while fuzzy logic is useful for decision making. The methods based on quantitative models are oriented to identification (parameter estimation), observers (state estimation) and parity space. Some heuristics results, obtained from the previous experience, can be used for diagnosing the origins of the failure or change, based on the dispersion of the characteristics.

Almost all CD solutions assume that the monitored system can be described, with sufficient precision, by a finite-dimensional linear model. In practice, if the system is more complex than the structure, described by a finite-dimensional model, the parameter estimates will still converge, but their values can be strongly dependent on the experimental conditions. The algorithms will not be able to separate the changes determined by the external conditions from those produced by the internal defect of the investigated system, so the classical tests will fail. The problems mentioned above point out the requirement of the robust CD algorithms, able to separate the changes determined by the external conditions from the changes of the internal dynamics of the system.

The first generation of CD algorithms is based on strong hypotheses, or strong assumptions, which are difficult to verify in practice. So, a second generation of solutions was required, insensitive to the uncertainty of the system's dynamics, to the operating environment, and to large noise, statistically unknown. In our opinion, among the central problems to be addressed in the CD area refer to robustness, sensitivity and versatility. The lack of robustness of the classical algorithms concerns the failure of the detection, if one or more of the hypotheses assumed during the design are not verified in practice. The sensitivity concerns the ability of the algorithm to detect the change, even if there are small scale incipient changes. Finally, the versatility concerns the ability of the methods and techniques to solve more CD problems, using the same set of algorithms.

To solve the vibration monitoring problem different techniques have been developed, one can be mentioned: analysis of overall vibration level, frequency spectrum, envelope spectrum, cepstrum analysis, etc. [5]. The success of vibration monitoring, in many practical cases, requires specialized functions and tools. Simple application of CD techniques on original mono- or multivariate vibration signals can assure the successful of monitoring. Sometimes, it is necessary that some signal pre- or postprocessing procedures to be applied, to emphasize and highlight the characteristics of the vibration signals making the object of the analysis. So, some signal processing techniques can be used in conjunction with CD techniques: independent component analysis (ICA), time-frequency analysis (TFA), energy distribution (ED) evaluation in time-frequency domain. These techniques are implemented in a software toolbox, Matlab VIBROTOOL Toolbox [6], built as a set of programs that compute specific parameters and solve specialized tasks for vibrating monitoring.

The CD problem can be solved by change point estimation (mean change), change detection using one and two model approach, with different distance measures and stoping rules

[7], multiple change detection [8], detection and diagnosis of model parameter and noise variance changes [9], for mono- and multivariable vibration signals. Some algorithms, making the object of [10] and [11] in CD, represented the starting points in developing of these algorithms. The analysis of the behavior of the vibration signals reveals that most of the changes that occur are either changes in the mean level, variance, or changes in spectral characteristics.

The toolbox is used in the framework offered by an experimental model, VIBROCHANGE, for vibrational processes analysis using advanced measuring and signal analysis techniques [12]. The main modules involved in VIBROCHANGE include:

- VIBROSIG - vibration and other signals measurement module.
- VIBROTOOL - Matlab Toolbox for change detection and diagnosis, with modules dedicated to change detection and segmentation problem solving, among others.
- VIBROMOD - hardware module for change detection and diagnosis implementing some components of the VIBROTOOL module, for on-line analysis and monitoring of vibration processes [13].

For laboratory condition working, a generator of vibrations in controlled operation mode, for different electro-mechanical processes, VIBROGEN, has been developed. It includes an electrical motor, as well as bearings and other gearboxes, to emulate an industrial process. The system is an open one and different incipient faults can be generated.

V. Signal Processing in Vibration Monitoring

The success of vibration monitoring requires specialized functions and tools to compute specific parameters and solve specialized tasks for change detection using classical and recent techniques.

Sometimes, only simple application of CD techniques on original mono- or multivariate vibration signals can assure the successful of monitoring. Frequently, it is necessary to apply some signal pre- or postprocessing procedures, to emphasize and highlight the characteristics of the vibration signals making the object of the analysis. So, some signal processing techniques can be used in conjunction with CD techniques: independent component analysis (ICA), time-frequency analysis (TFA), energy distribution (ED) evaluation in time-frequency domain, etc. These techniques are briefly described in in the following.

A. Change Detection - CD

We present here only the framework in which the CD problem will be solved in the case studies presented in Section VII, using the Maximum A posteriori Probability (MAP) estimator [8].

CD allows for a first detection of changes in the original vibration signals, or in other signals, resulting after a possible preprocessing of these. A frequently used model, in this case, is a linear regression model with piecewise constant parameters [8],

$$y_t = \phi_t^T \theta(i) + e_t, \;\; E(e_t^2) = R_t, \tag{1}$$

where y_t is the observed signal, $\theta(i)$ is the d-dimensional parameter vector in data stationary segment i, ϕ_t is the regressor; the noise e_t is assumed to be Gaussian with variance R_t.

The used model is referred to as changing regression, because it changes between regression models. Its important feature is that the jumps divide the vibration signals into a number of independent segments, since the parameter vectors in different segments are independent. Some important model derived from the model, where ϕ_t has different expressions [8]. In this framework, the problem of segmentation between "homogenous" parts of the data arises more or less explicitly.

B. Independent Component Analysis - ICA

Independent Component Analysis (ICA) is closely related to the blind source separation (BSS) [14], offering new solutions for vibration and noise analysis [15]. The use of BSS techniques in conjunction with other techniques, such as CD and TFA, proved very useful in vibration monitoring. So, it is offered the possibility to translate the CD problem from the original space of the measurements to the space of the independent components (sources). The reduced number of components, in this case, will simplify the monitoring problem and the CD methods will be applied only for scalar signals; BSS also provides a mixing model of the independent sources, that point out how the source changes are reflected in the original vibration signals, for diagnosis purposes. When it comes to deal with mechanical signals, which are typically characterized by an excessive complexity, BSS faces a number of difficulties which seriously hinder its feasibility [15].

One of the frequently used model for BSS, assumes the existence of n independent signals $s_1(t), \ldots, s_n(t)$ and the observation of as many mixtures $x_1(t), \ldots, x_n(t)$, these mixtures being linear and instantaneous, i.e.

$$x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t) + n_i(t) \tag{2}$$

for each $i = 1, n$, and compactly represented by the mixing equation

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{3}$$

where $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^T$ is an $n \times 1$ column vector containing the source signals, while vector $\mathbf{x}(t)$ contains the n observed signals and the square $n \times n$ "mixing matrix" $\mathbf{A}$ contains the mixture coefficients.

The BSS objective is to recover the source vector $\mathbf{s}(t)$ using only the observed data $\mathbf{x}(t)$, the assumption of independence between the entries of the input vector $\mathbf{s}(t)$ and possible some

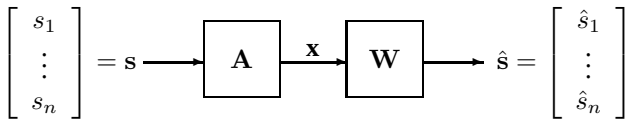


Fig. 1. Signal mixing and separating in BSS.

a priori information about the probability distribution of the inputs. It can be formulated as the computation of an $n \times n$ "separating matrix" $\mathbf{W}$ whose output $\hat{\mathbf{s}}(t)$ is an estimate of the vector $\mathbf{s}(t)$ of the source signals, and has the form:

$$\hat{\mathbf{s}}(t) = \mathbf{W}\mathbf{x}(t) \tag{4}$$

in the case of an instantaneous mixture (see Fig. 1).

For temporal coherent signals, the BSS problem can be solved using second and higher order statistics, the well known algorithms being SOBI (Second Order Blind Identification)) [16], and JADE (Joint Approximate Diagonalization of Eigenmatrices) [17], among others.

C. Time-Frequency Analyis - TFA

The analysis, processing, and parameter estimation of vibration signals whose spectral content changes in time are crucial in many CD applications. In this case, TFA can be of great interest, specially when the signal models are unavailable. In those cases, the time or the frequency domain descriptions of a signal alone cannot provide comprehensive information for change detection. The time domain lacks the frequency description of the signals. The TFA provides a proper description of the spectral content changes as a function of time.

The time-frequency representations (TFRs) can be classified according to the analysis approaches [18]. In the first category, the signal is represented by time-frequency (TF) functions derived from translating, modulating and scaling a basis function having a definite time and frequency localization. For a signal, $x(t)$, the TFR is given by

$$TF_x(t,\omega) = \int_{-\infty}^{+\infty} x(\tau)\phi^*_{t,\omega}(\tau)d\tau = < x, \phi_t, \omega >, \tag{5}$$

where $\phi_{t,\omega}$ represents the basis functions and $*$ represents the complex conjugate. The basis functions are assumed to be square integrable, i.e., they have finite energy. Short-time Fourier transform (STFT) [19], wavelets [20], [21], and matching pursuit algorithms [20], [22], are typical examples in this category.

The second category of time-frequency distributions (TFD), known as Cohen's shift invariant class distributions [3], characterizes the TFR by a kernel function. TABLE I gives the kernels used for main Cohen's class time-frequency distributions.

Some remarks on properties of the main Cohen's class time-frequency distributions from TABLE I could be made.

TABLE I. KERNELS USED FOR MAIN COHEN'S CLASS TIME-FREQUENCY DISTRIBUTIONS

Name	Kernel $\phi(\theta,\tau)$
SP	$\int h^*(u-\frac{1}{2}\tau)\exp^{-j\theta u}h(u+\frac{1}{2}\tau)du$
WVD	1
CWD	$\exp^{-\theta^2 t^2/\sigma^2}$
RID	2d Low pass filter in θ, τ space

The spectrogram (SP), suffers from the undesirable trade-off between the resolution and frequency resolution. On the other hand, the Wigner-Ville distribution (WVD) has a high time-frequency resolution, but is known to suffer from the presence of cross-terms. The Choi-Williams distribution (CWD) overcomes the WVD limitation suppressing to a large extent the cross-term interference, but some time-frequency resolution is lost. The last distribution belongs to the so-called Reduced Interference Distribution (RID), and also belongs to the Cohen's class, being an extension of the WVD.

Even if all TFDs tend to the same goal, each representation has to be interpreted differently, according to its own properties. For example, some of them present important interference terms, other are only positive, other are perfectly localized on particular signals, etc. The extraction of information has to be done with care, from the knowledge of these properties. We need a distribution that can reveal the features of the signal as clearly as possible without any "ghost" component and to apply a TFD that can get rid of the cross-terms while preserving a high time-frequency resolution.

D. Energy Distribution - ED

One of the simplest feature based signal processing procedures in TFA is via energy distribution. The idea is to analyze the distribution of energy at certain time instant or certain frequency band or more generally, in some particular time and frequency region. Such analysis is capable of revealing more information about a particular phenomenon [2], [18].

Once the local frequency content has been obtained, using TFA, an entropy measure can be evaluated for extracting the information containing in a given position of $t = n$. The Rényi entropy measures class [23], [24], with some significant contributions [25], offers new measures for estimating signal information and complexity in the time-frequency plane.

For a generic time-frequency distribution, $P_x(n,k)$, the Rényi entropy measure has the following form:

$$R_\alpha = \frac{1}{1-\alpha}\log_2\left(\sum_n\sum_k P_x^\alpha(n,k)\right) \tag{6}$$

where n is the temporal discrete variable and k the frequency discrete variable, with $\alpha \geq 2$ being values recommended for time-frequency distribution measures [25]. The normalized Rényi entropy measures, with the normalization done in various ways, leads to a variety of possible measure definitions [2], [25]. Eisberg and Resnik [26], assimilate the time-frequency

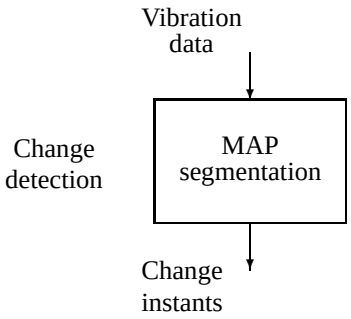


Fig. 2. First approach in machine health monitoring

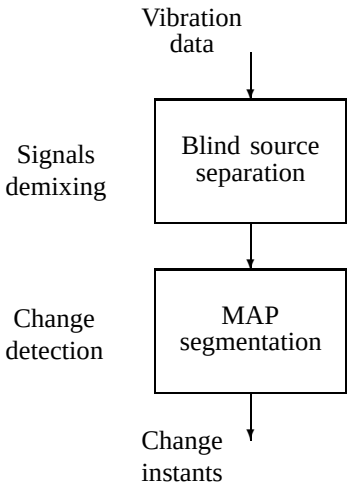


Fig. 3. Second approach in machine health monitoring

distributions at a given instant $t = n$ to a wave function and for $\alpha = 3$, resulting

$$R_3 = -\frac{1}{2}\log_2\left(\sum_n\sum_k P_x^3(n,k)\right) \tag{7}$$

The normalizing stage affects exclusively to index k, when the operation is restricted to a single position n to satisfy the condition $\sum_k P_x(n,k) = 1$ in such position.

The measure (7) can be rewritten for a given n as follows:

$$R_3(n) = -\frac{1}{2}\log_2\left(\sum_k P_x^3(n,k)\right) \tag{8}$$

Empirically the normalization proposed in [26] had shown to be most suitable for an application in vibration signal analysis. The values of $R_3(n)$ depend upon the size N of the window and it can be shown that they are within the interval $0 \leq R_3(n) \leq \log_2 N$. Hence, the measure can be normalized by applying $\hat{R}_3(n) = R_3(n)/\log_2 N$.

VI. General Approach for Change Detection

The signal processing techniques mentioned above can be used in different combinations to solve the problem of machine health monitoring. Three main approaches are discussed in the following.

A first approach simply consists of original signal segmentation (see Fig. 2), resulting the change points in vibration signal dynamics. The MAP algorithm [8], is one algorithm which can be used in this case, with good results for mono- and multivariate signals. Some experimental results, using this approach, in simulation and with real data, are presented in [8], [27].

A second approach (see Fig. 3) makes use of change detection of the signals resulted after blind source separation of independent vibration sources, starting from the original vibration signals. In this case, the problem is transferred from the original space of the measurements to the space of independent sources, where the reduced number of components will simplify the health monitoring problem, and the change detection methods will be applied for scalar signals. The assessment of the approach on a real machine is presented in [7].

The third approach, considered as a complex and general approach, practically includes all the signal processing techniques discussed above and is given in Fig. 4.

The approach makes use of time-frequency information content, the short-term time-frequency Rényi entropy, and a segmentation algorithm, based on MAP estimator. The segmentation algorithm operates on Rényi entropy, as a new space of decision. The procedure can be applied on the original vibration signals, or on the independent vibration sources resulted for these, after blind source separation. This approach enables more robust change detection in vibration signals. The application of the presented approach offers a simpler analysis and interpretation of the vibration signals behavior, providing new physical insight into vibrational processes. Same experimental results in simulation and with real data are given in [28], [29], [30].

VII. Case Studies

This section presents some experimental results obtained in two case study having as object fault detection in rolling element bearings (REB) and in a rotating machine, a pump, using the framework described in the previous sections of the paper.

A. Fault Detection in Rolling Elements Bearings

1) Test Data: The performed experiments use a data set from [31], with three faults having different locations: $F1$ (Inner race), $F2$ (Ball) and $F3$ (Outer race), and four sizes of the faults; $F0$ denotes no faults; only the data for the first case (06HH) have been used (see TABLE II).

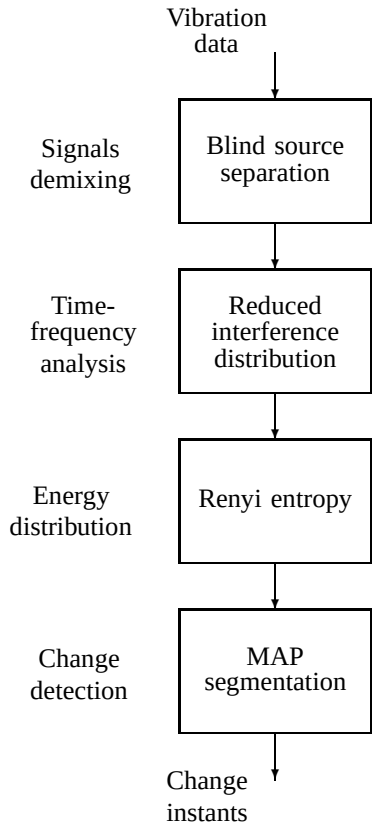


Fig. 4. Third approach in machine health monitoring

TABLE II. 1ST DATA TEST SET (6203 BEARING TYPE)

Fault size	$F0$ Free	$F1$ Inn. Race	$F2$ Ball	$F3$ Outer Race
0.000"	$y_0(t)$	-	-	-
0.007"	-	$y_1(t)$	$y_2(t)$	$y_3(t)$
0.014"	-	$y_4(t)$	$y_5(t)$	$y_6(t)$
0.021"	-	$y_7(t)$	$y_8(t)$	$y_9(t)$
0.028"	-	$y_{10}(t)$	$y_{11}(t)$	-

The signal $y_0(t)$ contains 4,096 samples recorded during normal conditions operating, while signals $y_i(t)$, $i = 1, \dots 11$ indicate files/vectors, containing each 4,096 samples, for the cases with faults; the sampling rate was of 12,000 samples/s.

2) Preliminary Analysis: For the signals mentioned above, some statistical features in time domain [32], have been computed, and are given in TABLE III, offering a general view of the signal characteristics.

The signals, making the object of the analysis, are simultaneously characterized in time and frequency domain using their mean localizations and dispersions. So, the averaged time and the time spreading, as well as the averaged frequency and the frequency spreading [33], are given in TABLE IV for signals analyzed.

3) Algorithm Description: The model used in the case study is a linear regression model with piecewise constant parameters (1).

TABLE III. TIME-DOMAIN STATISTICAL FEATURES OF THE SIGNALS $y_0(t), y_1(t), \dots, y_{11}(t)$ IN TIME DOMAIN

Signal	RMS	Mean	Var.	Cres. fact.	Skew.	Kurt.
$y_0(t)$	0.999	-0.002	0.998	3.796	-0.094	2.890
$y_1(t)$	0.992	0.007	0.985	5.145	0.124	5.456
$y_2(t)$	1.007	0.021	1.014	3.720	0.003	2.997
$y_3(t)$	0.997	0.016	0.995	5.189	0.088	7.698
$y_4(t)$	0.997	-0.001	0.995	4.016	0.067	4.281
$y_5(t)$	1.013	0.013	1.027	5.299	0.012	7.032
$y_6(t)$	0.987	0.078	0.974	9.747	-0.144	22.505
$y_7(t)$	0.724	0.001	0.525	6.937	-0.066	5.775
$y_8(t)$	0.978	0.046	0.958	3.779	0.023	2.982
$y_9(t)$	1.018	0.011	1.037	6.495	0.315	6.868
$y_{10}(t)$	0.981	0.019	0.963	4.378	0.043	3.457
$y_{11}(t)$	0.955	0.002	0.913	9.992	-0.086	21.255

TABLE IV. TIME-FREQUENCY STATISTICAL FEATURES OF THE SIGNALS $y_0(t), y_1(t), \dots, y_{11}(t)$

Signal	Aver. time	Time spread	Aver. freq.	Freq. spread
$y_0(t)$	2.104e+003	4.251e+003	-8.197e-009	0.287
$y_1(t)$	2.032e+003	4.155e+003	-2.359e-008	0.850
$y_2(t)$	2.026e+003	4.103e+003	-1.035e-006	0.906
$y_3(t)$	2.090e+003	4.167e+003	-2.206e-008	0.969
$y_4(t)$	1.944e+003	4.157e+003	-5.457e-009	0.804
$y_5(t)$	2.082e+003	4.247e+003	-3.880e-008	0.983
$y_6(t)$	1.954e+003	4.099e+003	-1.229e-008	0.920
$y_7(t)$	1.993e+003	4.843e+003	-1.134e-008	0.820
$y_8(t)$	2.057e+003	4.187e+003	-1.800e-007	0.968
$y_9(t)$	2.054e+003	4.273e+003	-1.604e-007	0.857
$y_{10}(t)$	2.006e+003	4.184e+003	-1.435e-007	0.909
$y_{11}(t)$	2.085e+003	4.081e+003	-9.584e-010	0.911

To solve the segmentation problem, all possible segmentation k^n are considered, estimate one linear regression model in each segment, and then choose the particular k^n that minimizes an optimality criteria of the form:

$$\widehat{k^n} = \arg \min_{n \geq 1, 0 < k_1 < \dots < k_n = N} V(k^n) \tag{9}$$

For the measurements in a i-th segment, $y_{k_{i-1}+1}, \dots y_{k_i} = y_{k_{i-1}+1}^{k_i}$, results the least square estimate and its covariance matrix:

$$\hat{\theta}(i) = P(i) \sum_{t=k_{i-1}+1}^{k_i} \phi_t R_t^{-1} y_t, \tag{10}$$

$$P(i) = \left(\sum_{t=k_{i-1}+1}^{k_i} \phi_t R_t^{-1} \phi_t^T \right)^{-1}. \tag{11}$$

The following quantities are used in optimal segmentation algorithm:

$$V(i) = \sum_{t=k_{i-1}+1}^{k_i} (y_t - \phi_t^T \hat{\theta}(i))^T R_t^{-1} (y_t - \phi_t^T \hat{\theta}(i)) \tag{12}$$

$$D(i) = -\log \det P(i) \tag{13}$$

$$N(i) = k_i - k_{i-1} \tag{14}$$

where $V(i)$ - the sum of squared residuals, $D(i)$ - $-\log\det$ of the covariance matrix $P(i)$ and $N(i)$ - the number of data in each i segment, and represent sufficient statistics for each segment. The data and quantities used in segmentation k^n, having $n-1$ degrees of freedom are given in TABLE V.

TABLE V. DATA AND QUANTITIES USED IN OPTIMAL SEGMENTATION PROCEDURE

Data	$y_1, y_2, \ldots, y_{k_1}$	...	$y_{k_{n-1}+1}, \ldots, y_{k_n}$
Segment	Segment 1	...	Segment n
LS est.	$\hat{\theta}(1), P(1)$	...	$\hat{\theta}(n), P(n)$
Statistics	$V(1), D(1), N(1)$	...	$V(n), D(n), N(n)$

To solve the optimal segmentation procedure, different types of optimality criteria have been proposed [11]. In the following we use MAP criterium [8]. The number of segmentations k^n is 2^N (can be a change or no change at each time instant), and this put problems concerning the dimensionality.

The conceptual description MAP estimator, for the data and quantities given in TABLE IV, is given in Fig. 5, for three different assumptions on noise scaling: (i) known $\lambda(i) = \lambda_0$, (ii) unknown but constant $\lambda(i) = \lambda$ and (iii) unknown and changing $\lambda(i)$, where q is the change probability at each time instants ($0 < q < 1$).

Data: Vibration signal $y_t, \quad t = 1 \ldots N$

Step 1: Examine every possible segmentation, parameterized in the number of jumps n and jump times k^n, separately.

Step 2: For each segmentation, compute the best models in each segment parameterized in the least square estimates $\hat{\theta}(i)$ and their covariance matrices $P(i)$.

Step 3: Compute in each segment:

$$\begin{aligned} V(i) &= \textstyle\sum_{t=k_{i-1}+1}^{k_i} (y_t - \phi_t^T \hat{\theta}(i))^T R_t^{-1} (y_t - \phi_t^T \hat{\theta}(i)) \\ D(i) &= -\log\det P(i) \\ N(i) &= k_i - k_{i-1} \end{aligned}$$

Step 4: MAP estimate, $\widehat{k^n}$, for the three different assumptions on noise scaling

(i) known $\lambda(i) = \lambda_0$,

$$\widehat{k^n} = \arg\min_{k^n, n} \textstyle\sum_{i=1}^{n} (D(i) + V(i)) + 2n \log \frac{1-q}{q}$$

(ii) unknown but constant $\lambda(i) = \lambda$,

$$\widehat{k^n} = \arg\min_{k^n, n} \textstyle\sum_{i=1}^{n} D(i) + (Np - nd - 2) \times \\ \times \log \sum_{i=1}^{n} \frac{V(i)}{Np - nd - 4} + 2n \log \frac{1-q}{q}$$

(iii) unknown and changing $\lambda(i)$,

$$\widehat{k^n} = \arg\min_{k^n, n} \textstyle\sum_{i=1}^{n} (D(i) + (N(i)p - d - 2) \times \\ \times \log \frac{V(i)}{N(i)p - d - 4}) + 2n \log \frac{1-q}{q}$$

Results : Number n and locations k_i , $k^n = k_1, k_2, \ldots, k_n$

Fig. 5. MAP segmentation algorithm.

In a practical problem, only one of the equations from **Step 4** (see Fig. 5) is evaluated, according with the assumption on noise scaling of the procedure.

For the exact likelihood evaluation, can be implemented recursive local search techniques and numerical searches based on dynamic programming or MCMC (Markov Chain Monte Carlo) techniques [11], [8].

Starting from the optimal segmentation results it is possible to analyze the data resulted for each stationary data segment to locate and diagnose the produced fault or change in the REB: outer race, inner race, bearing cage, ball (roller), according with the frequency area where it was produced.

4) Multiple Fault Detection: Started from the data given in TABLE II data sequences with multiple faults have been generated, for 3 types of events: inner race faults, ball faults and outer race faults, with different fault size: 0.007", 0.014", 0.021", 0.028", for the first two cases, and 0.007", 0.014", 0.021" for the third case. The following data sets have been used in the analysis, for fault detection:

$$\begin{aligned} s_1(t) &= [y_0(t), y_1(t), y_4(t), y_7(t), y_{10}(t)] \\ s_2(t) &= [y_0(t), y_2(t), y_5(t), y_8(t), y_{11}(t)] \\ s_3(t) &= [y_0(t), y_3(t), y_6(t), y_9(t)] \end{aligned}$$

resulting data sequences of 20480 values for signals $s_1(t), s_2(t)$ and 16384 for signal $s_3(t)$. The real faults instants were 4097, 8193, 12288 and 16384. These data sets offer the possibility to fault detection of a graduate size of fault, for the cases mentioned above.

The experimental results refer to the signals $s_1(t), s_2(t), s_3(t)$ and the segmenting algorithm presented above with unknown and constant noise scaling, and MCMC algorithm [8], with a value of jump probability, $q = 0.3$ and appropriate design parameters in search scheme, for different model order, na. The fault instants detected for different model orders na are presented in TABLE VI, TABLE VII and TABLE VIII for $s_1(t)$, $s_2(t)$ and $s_3(t)$, respectively.

The signal $s_1(t)$, making the object of the analysis, and the estimated multiple fault times for the inner race, $na = 20$ and $q = 0.3$, are presented in Fig. 6, while the signal $s_2(t)$ and the estimated multiple fault times for ball, $na = 20$ and $q = 0.3$ are given in Fig. 7. The signal $s_3(t)$ and the estimated multiple fault times for the outer race, $na = 60$ and $q = 0.3$ make the object of Fig. 8.

TABLE VI. FAULT DETECTION IN SIGNAL $s_1(t)$ USING DIFFERENT MODEL ORDER

Model order	Fault detection instants
$na = 10$	4096, 8687, 9501, 10684, 11322, 11500, 12570, 12627, 12967, 13068, 13961, 14527, 14627, 14777, 15964, 16384.
$na = 15$	4096, 8687, 9502, 10684, 11501, 12570, 14777, 16384.
$na = 20$	4096, 8195, 8687, 11502, 13026, 16384.

The changes in signals $s_1(t)$, $s_2(t)$ and $s_3(t)$, resulted after the data concatenation, are gradual, whose effect may

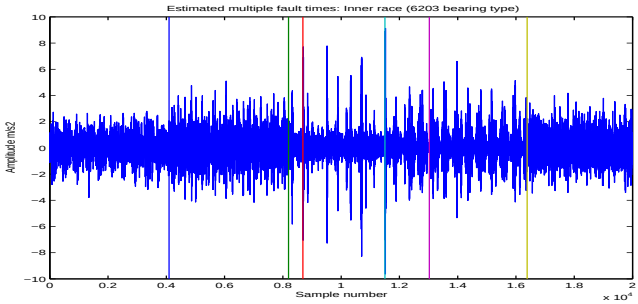


Fig. 6. The signal $s_1(t)$ and estimated multiple fault times for inner race, $na = 20$, $q = 0.3$.

TABLE VII. FAULT DETECTION IN SIGNAL $s_2(t)$ USING DIFFERENT MODEL ORDER

Model order	Fault detection instants
$na = 10$	4096, 8191, 8497, 8614, 9305, 9929, 11946, 16385, 16711, 16901, 18065, 18129.
$na = 15$	4096, 8190, 11946, 16385, 16719, 18108, 18128.
$na = 20$	4096, 8190, 11945, 16385, 16751, 18233.

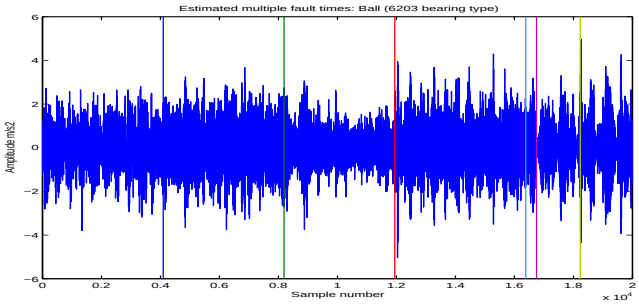


Fig. 7. The signal $s_2(t)$ and estimated multiple fault times for the ball, $na = 20$, $q = 0.3$.

TABLE VIII. FAULT DETECTION IN SIGNAL $s_3(t)$ USING DIFFERENT MODEL ORDER

Model order	Fault detection instants
$na = 10$	4096, 4383, 7081, 7170, 7897, 7950, 8192, 12298, 12367, 12480, 12982, 13151, 13260, 13407, 13596, 14042, 14179, 14378, 14489, 14668, 14823, 15169, 15271, 15575, 15605, 16050, 16229.
$na = 15$	4096, 8192, 12296, 12368, 12479, 12669, 12813, 13261, 13455, 13596, 14042, 14173, 14378, 15015, 15164, 15271, 15469, 15605, 16051, 16346.
$na = 20$	4096, 8192, 12293, 12367, 12479, 12669, 12813, 13261, 13460, 13594, 14042, 14189, 14378, 15271, 15473, 15604, 16051.
$na = 60$	4096, 8198, 12287, 12352, 14057.

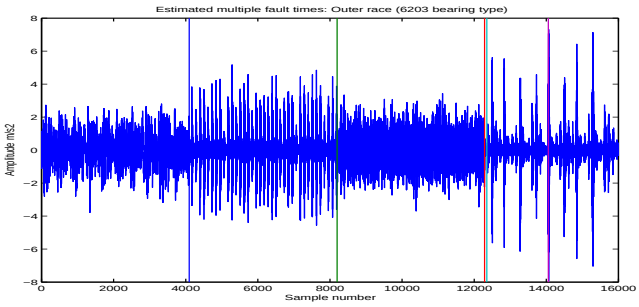


Fig. 8. The signal $s_3(t)$ and estimated multiple fault times for outer race, $na = 60$, $q = 0.3$.

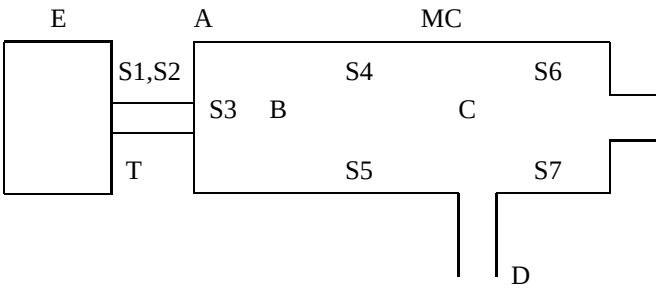


Fig. 9. Schematically multichannel measurement

increase, producing new changes in the signal dynamics that can be detected by the algorithm. The further deterioration of the rolling element bearing during operating produces new fault instants, different from 4096, 8192, 12288 and 16384 instants. According with data from TABLE VI, TABLE VII and TABLE VIII one can notice that in all the cases the main faults are detected. Also, it can be noted that for the models with high order ($na = 20$, $na = 20$ and $na = 60$, respectively), only the main faults are detected at instants 4096, 8192, 12288 and 16384 or near instants. The models, of high order, can increase the robustness of the optimal segmentation algorithm to gradual, or small changes in signal dynamics. Different values of q offer similar results, but a higher order of the model leads to a better fault detection, the model being more able to approximate the signal dynamics.

B. Industrial Pump Monitoring

The machine under investigation is an industrial pump. The used data set consists of multichannel measurements for 7 channels repeated for two identical machines: the first is virtually fault free and the second shows a progressed pitting in both gears [34]. The data were selected from the high-frequency measurements, digitized at 12800 Hz, a data segment of 4096 values, 2048 from the fault free machine, and last 2048 from the machine with a progressed pitting in gears, both for minimum load. The data have been low-pass filtered to 5000 Hz.

A scheme of the machine with its components and sensor position is given in Fig. 9, with the following legend:

E	=	electromotor
A	=	incoming shaft (driving shaft)
MC	=	machine casing
T	=	tachometer
B	=	first delay (gear-combination)
C	=	second delay (gear-combination)
D	=	outgoing shaft (to vane in water)
S1-S7	=	position of sensors 1,7

The rotating speed of the driving shaft is measured with a tachometer. This measurement is done synchronously with

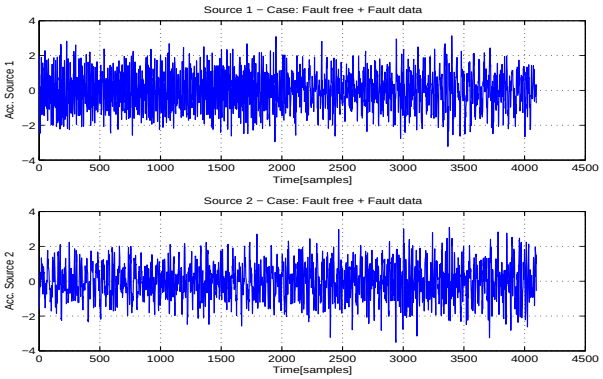


Fig. 10. Independent vibration sources in normal operating and fault conditions of the pump

7 accelerometers used in the following manner: sensor S1,S2 are radially mounted near the driving shaft, with an angle of 90 degrees between them, sensor S3 is axially mounted near the driving shaft, and sensor S4-S7 are radially mounted on different parts of the machine.

The data represented the object of the analysis in [7], where the blind source separation (BSS) and change detection in source signals, according with the approach presented in Fig. 3. The case study, making the object of this section, used the general approach, given in Fig. 4, for the data set mentioned above. It includes vibration signal demixing, time-frequency analysis, energy distribution evaluation using short-term Rényi entropy, and its segmentation, based MAP estimator. The segmentation algorithm operates on Rényi entropy, as a new space of decision. We discuss in the following this approach and present the experimental results.

1) Blind Source Separation: The acceleration measurements for 7 channels and 4098 values, from the fault free machine and progressed pitting in gears machine, represented the input data for SOBI algorithm [16], when 2 independent vibration sources and an instantaneous mixture model have been considered. The number of the sources resulted via eigendecomposition of the sample covariance matrix [35]. The independent vibration sources are presented in Fig. 10.

2) Time-Frequency Rényi Entropy: Fig. 11 shows the reduced interference distribution (RID) [36], of S1 source, computed with a kernel based on the Hanning window [33]. In Fig. 11 at linear scale, it can be noted a change in the spectral content of the source, in the second part of the signal.

Results of the TFD analysis for S2 source are presented in Fig. 12 for RID, with a same Hanning window. Similar conclusions, as in the previous analyzed case, concerning TFD properties, could be established. From Fig. 12 it can be noted a reduced change in the spectral content of the source, in the second part of the signal, in comparison with the S1 source.

A first conclusion, in this stage of time-frequency analysis, could be that S1 source has been induced by the fault in pump gears, but because the source separation is not perfect, due to

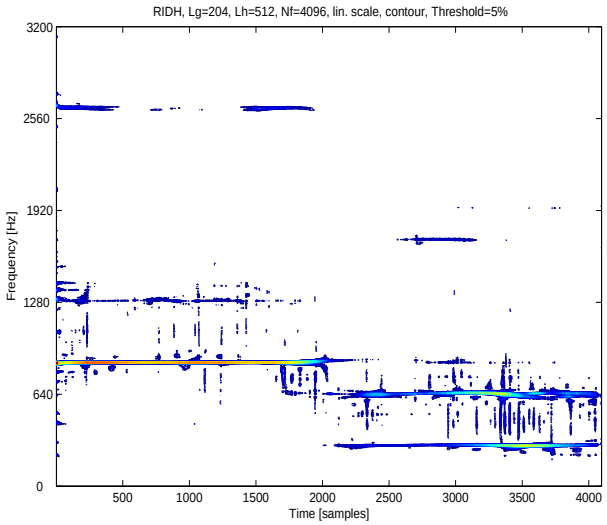


Fig. 11. Reduced interference distribution for vibration source S1 in normal and fault operating conditions

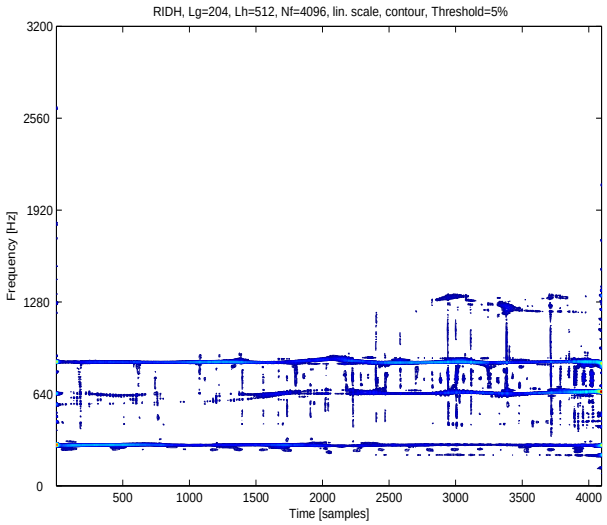


Fig. 12. Reduced interference distribution for vibration source S2 in normal and fault operating conditions

a possible lack of the BSS method robustness, the real change can also induced in other sources, in our case in the S2 source. The differences between the changes in spectral content in both sources point out this fact.

To evaluate the TFD resulted for S1 source we present in Fig. 13 the short-term Rényi entropy as measure of time-frequency distribution, computed for RID. It was used a sliding window of $N = 64$ values and a constant bias to be added to signal of 1.

For S2 source, we present in Fig. 14 the short-term Rényi entropy, as measure of time-frequency distribution, computed for RID, with the same values for the sliding window and constant bias added to signal.

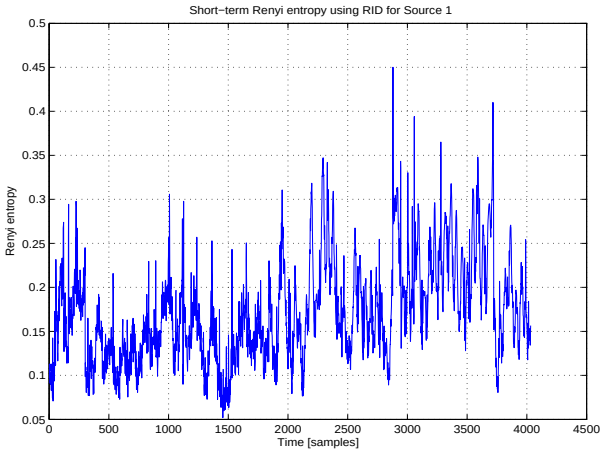


Fig. 13. Short-term Renyi entropy using RID for source S1

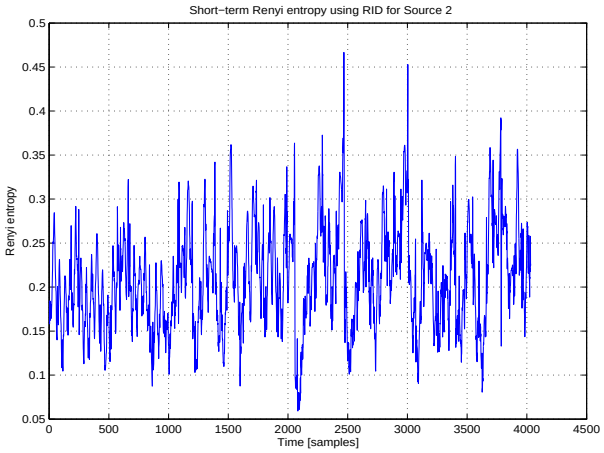


Fig. 14. Short-term Renyi entropy using RID for source S2

3) MAP Segmentation of Rényi Entropy: Visual inspection for the Rényi entropy of both sources, shows that the onset time is clearly visible as a change in energy and frequency content. Our experience is that, for this problem, as for many other signal processing ones, a piecewise constant model (1), could lead to a satisfactory trade-off between complexity and efficiency of the corresponding algorithms for the off-line estimation of the change time. The segmentation procedure has been performed using an autoregressive model (AR) of order 1, the unknown and constant noise scaling assumption and MCMC algorithm.

The parameter and variance estimates resulted in MAP segmentation are presented in Fig. 15 and Fig. 16 for Rényi entropies, obtained for S1 and S2 sources, respectively.

The variance traces of the piecewise constant model show, for both sources, significant jumps in the second part of the signals, and that a main distinct rupture event occurred. The proposed procedure assures more robust change detection in vibration signal analysis, than in the case of change detection in the estimated sources in time domain, see [7].

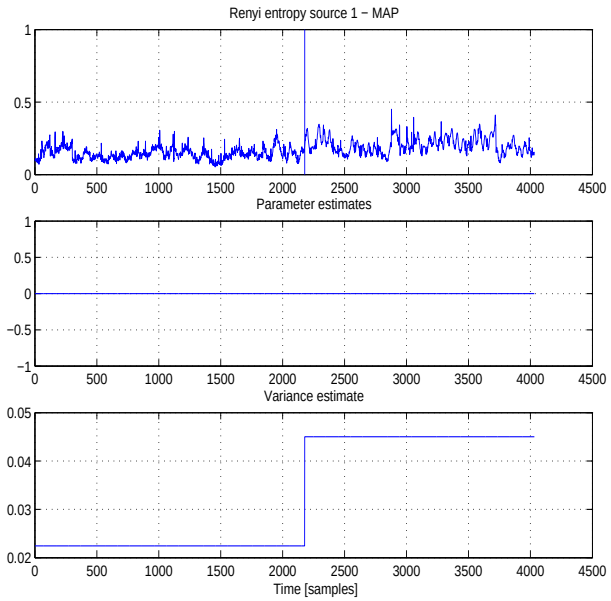


Fig. 15. MAP segmentation of short-term Renyi entropy for source S1

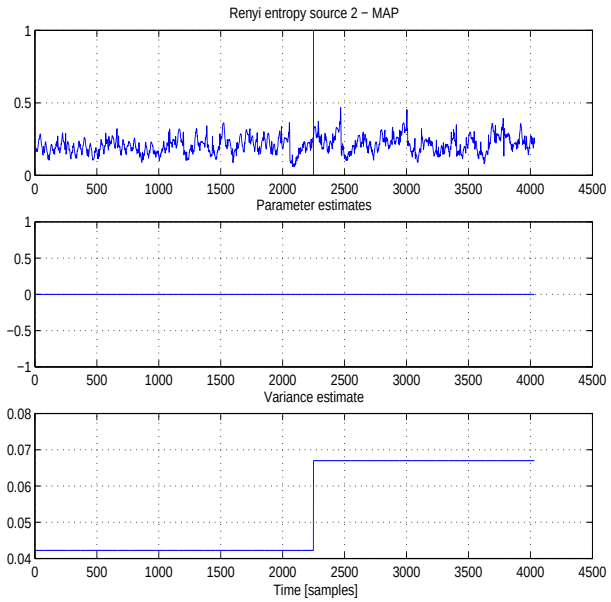


Fig. 16. MAP segmentation of short-term Renyi entropy for source S2

VIII. Conclusions

The paper considers the problem of change detection in vibration signals, with application in predictive maintenance of rotating machines, integrating some signal processing techniques, mainly independent component analysis, time-frequency analysis, energy distribution evaluation in time-frequency domain, and a change detection algorithm based on MAP estimator.

The case studies making the object of the paper prove the effectiveness of the proposed approach. The first case study,

having as subject detection of faults in REB, uses a segmentation algorithm based on MAP estimator, directly applied to vibration signals, while the second, for monitoring of an industrial pump, makes use of time-frequency Rényi entropy segmentation, applied to independent vibration sources of the pump.

The general approach offers new possibilities for more robust detection of changes in vibrating signals and assures proactive actions in vibration monitoring. It offers a simpler analysis and interpretation of the vibration signals behavior, providing new physical insight into vibration processes for predictive maintenance. It can also be used for other domains that require change detection and diagnosis, such as biomedical signal processing (EEG, EKG, and MEG), seismic signal analysis, infrastructure monitoring, speech analysis, communication systems, video surveillance, transportation systems, etc.

ACKNOWLEDGMENT

The paper was partially supported by 2019-2022 Romanian Core Program, Project RO-Smart Ageing, funded by Ministry of Research and Innovation.

REFERENCES

[1] Th. D. Popescu, D. Aiordachioaie and A. Culea-Florescu, "Vibration Analysis with Application in Predictive Maintenance of Rolling Element Bearing", *Proc. of International Conference on Emerging Networks and Systems Intelligence (EMERGING '2019)*, 22-26 September 22 - 26, 2019, Porto, Portugal.

[2] L. Stankovic, "A Measure of Some Time-Frequency Distributions Concentration", *Signal Processing*, pp. 621-631, 2001.

[3] L. Cohen, *Time-Frequency Distribution*, Prentice Hall, New York, 1995.

[4] S. Aviyente, "Information Processing on the Time-Frequency Plane", *Proc. IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP '04)*, pp. 617-620, 2004.

[5] S. .J. Lacey, "The Role of Vibration Monitoring in Predictive Maintenance", *FAG Technical Publication* Schaeffler Limited, UK, 2010.

[6] Th. D. Popescu and D. Aiordachioaie, "VIBROTOOL - Software Tool for Change Detection and Diagnosis in Vibration Signals", *Proc. of 59th IEEE International Midwest Symposium on Circuits and Systems (MWSCAS 2016)*, Abu Dhabi, United Arab Emirates, 16-19 October, pp. 640-643, 2016.

[7] Th. D. Popescu, "Blind Separation of Vibration Signals and Source Change Detection - Application to Machine Monitoring", *Applied Mathematical Modelling*, pp. 3408-3421, 2010.

[8] Th. D. Popescu, "Signal Segmentation using Changing Regression Models with Application in Seismic Engineering", *Digital Signal Processing*, pp. 14-26, 2014.

[9] Th. D. Popescu, "Detection and Diagnosis of Model Parameter and Noise Variance Changes with Application in Seismic Signal Processing", *Mechanical Systems and Signal Processing*, pp. 1598-1616, 2011.

[10] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes - Theory and Applications*, Prentice Hall, N.J., 1993.

[11] F. Gustafsson, *Adaptive Filtering and Change Detection*, Wiley, 2001.

[12] D. Aiordachioaie and Th. D. Popescu, "VIBROCHANGE - A Development System for Condition Monitoring Based on Advanced Techniques of Signal Processing", *The International Journal of Advanced Manufacturing Technology*, vol. 34, no. 11, pp. 3408-3421, 2019.

[13] D. Aiordăchioaie and Th. D. .Popescu, "VIBROMOD An Experimental Model for Change Detection and Diagnosis Problems", *Proc. The 14th IMEKO TC10 Workshop Technical Diagnostics, New Perspectives in Measurements, Tools and Techniques for systems reliability, maintainability and safety*, Milan, Italy, pp. 317-322, 2016.

[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley, 2001.

[15] J. Antoni, "Blind Separation of Vibration Components: Principles and Demonstrations", *Mechanical Systems and Signal Processing* pp. 1166-1180, 2005.

[16] A. Belouchrani, K. Abed Meraim, J. F. Cardoso, and E. Moulines, "A Blind Source Separation Technique using Second - Order Statistics", *IEEE Trans. Signal Processing*, pp. 434-444, 1977.

[17] J. F. Cardoso and A. Souloumiac, "Blind Beamforming for Non Gaussian Signals", *IEE Proceedings-F*, pp. 362-370, 1993.

[18] E. Sejdić, I. Djurović and J. Jiang, "Time-Frequency Feature Representation using Energy Concentration: An Overview of Recent Advances", *Digital Signal Processing*, pp. 153-183, 2009.

[19] K. Gröchenig, *Foundations of Time-Frequency Analysis*, Birkhäuser, 2001.

[20] S. G. Mallat, *A Wavelet Tour of Signal Processing*, Second Ed., Academic Press, 1999.

[21] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, 1992.

[22] S. G. Mallat and Z. Zhang, "Matching Pursuits with Time-Frequency Dictionaries", *IEEE Trans. Signal Process.* pp. 3397-3415, 1993.

[23] W. J. Williams, M. L. Brown and A. O. Hero, "Uncertainity, Information and Time-Frequency Distributions", *SPIE Adv. Signal Process. Algebra Arch. Imp.* pp. 144-156, 1991.

[24] T. H. Sang and W. J. Williams, "Rényi Information and Signal Dependent Optimal Kernel Design", *Proc. of the ICASSP* pp. 997-1000, 1995.

[25] P. Flandrin, R. G. Baraniuk and O. Michel, "Time-Frequency Complexity and Information", *Proc. of the ICASSP*, pp. 329-332, 1994.

[26] R. Eisberg and R. Resnick, *Quantum Physics*, Wiley, 1974.

[27] Th. D. Popescu and D. Aiordachioaie, "Fault Detection of Rolling Element Bearings using Optimal Segmentation of Vibrating Signals", *Mechanical Systems and Signal Processing* pp. 370-391, 2019.

[28] Th. D. Popescu and D. Aiordachioaie, "Signal Segmentation in Time-Frequency Plane using Rényi Entropy - Application in Seismic Signal Processing", *Proc. of The 2-nd IEEE International Conference on Control and Fault-Tolerant Systems (SysTol13)*, Nice, France, 9-11 October, pp. 312-317, 2013.

[29] Th. D. Popescu and B. Dumitrascu, "An Application of Rényi Entropy Segmentation in Fault Detection of Rotating Machinery", *Proc. of The 16th IEEE International Conference on Research and Education in Mechatronics (REM 2015)*, Bochum, Germany, 18-20 November, pp. 288-295, 2015.

[30] Th. D. Popescu and D. Aiordachioaie, "New Procedure for Change Detection Operating on Rényi Entropy with Application in Seismic Signals Processing", *Circuits, Systems, and Signal Processing*, pp. 3778-3798, 2017.

[31] * * * Case Western Reserve University Bearing Data Center, http://csegroups.case.edu/bearingdatacenter/home, 2017.

[32] T. R. Lin, K. Yu and J. Tan, *Condition Monitoring and Fault Diagnosis of Roller Element Bearing*, INTECH, pp. 39-75, 2017, http://www.intechopen.com/books/bearing-technology.

[33] P. Flandrin, *Temps-Fréquence*, Hermes, 1998.

[34] A. Ypma, *Learning Methods for Machine Vibration Analysis and Health Monitoring*, PhD Thesis, TU Delft, 2001.

[35] Y. Yin and P. Krishnaiah, "Methods for Detection of the Number of Signals", *IEEE Trans. on ASSP*, pp. 1533-1538, 1987.

[36] J. Jeong and W. J. Williams, "Kernel Design for Reduced Interference Distributions", *IEEE Transactions on Signal Processing*, pp. 402-4012, 1992.

A Low-Latency Power-Efficient Convolutional Neural Network Accelerator for Vision Processing Algorithms

Junghee Lee

School of Cybersecurity
Korea University
Seoul, Korea
Email: j_lee@korea.ac.kr

Chrysostomos Nicopoulos

Department of Electrical and Computer Engineering
University of Cyprus
Nicosia, Cyprus
Email: nicopoulos@ucy.ac.cy

***Abstract*—Deep Convolutional Neural Networks (CNN) are expanding their territory to many applications, including vision processing algorithms. This is because CNNs achieve higher accuracy compared to traditional signal processing algorithms. For *real-time* vision processing, however, their high demand for computational power and data movement limits their applicability to battery-powered devices. For such applications that require both real-time processing and power efficiency, hardware accelerators are inevitable in meeting the requirements. Recent CNN frameworks, such as SqueezeNet and GoogLeNet, necessitate a redesign of hardware accelerators, because their irregular architectures cannot be supported efficiently by traditional hardware accelerators. In this paper, we propose a novel hardware accelerator for advanced CNNs aimed at realizing real-time vision processing with high accuracy. The proposed design employs data-driven scheduling that enables support for irregular CNN architectures without run-time reconfiguration, and it offers high scalability through its modular design concept. Specifically, the design's on-chip memory management and on-chip communication fabric are tailored to CNNs. As a result, the new accelerator completes all layers of SqueezeNet and GoogLeNet in 14.30 ms and 27.12 ms at 2.47 W and 2.51 W, respectively, with 64 processing elements. The performance offered by the proposed accelerator is comparable to high-performance FPGA-based approaches (that achieve 1.06 to 262.9 ms at 25 to 58 W), albeit with significantly lower power consumption. If the hardware budget allows, these latencies can be further reduced to 6.71 ms and 11.70 ms, respectively, with 256 processing elements. In comparison, the latency reported by existing architectures executing large-scale deep CNNs ranges from 115.3 ms to 4309.5 ms.**

Keywords–Convolutional neural network; Hardware accelerator; On-chip memory optimization; On-chip communication

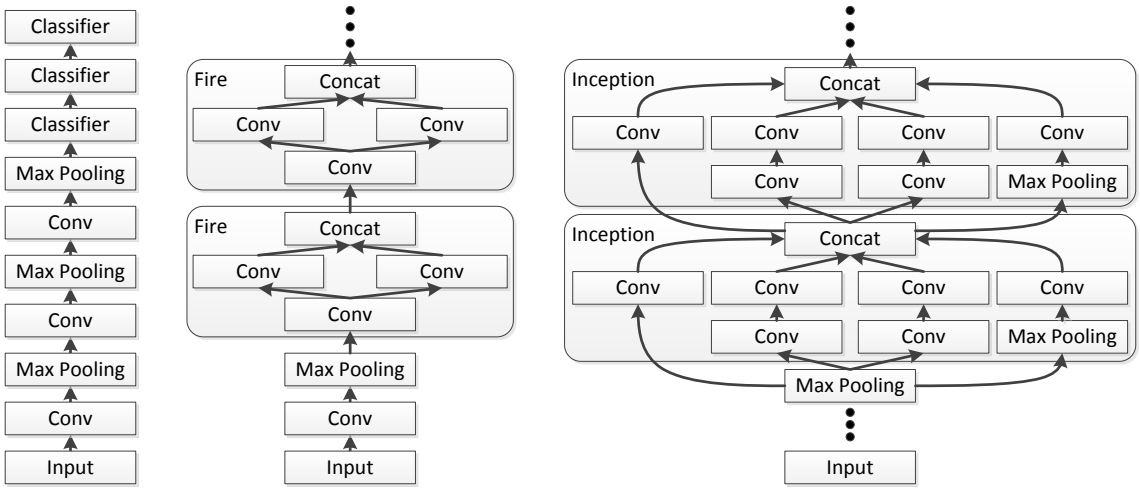
I. INTRODUCTION

As unmanned vehicles and robotics keep evolving, there is a growing demand for power-efficient real-time vision processing. While deep Convolutional Neural Networks (CNN) offer high accuracy and are applicable to various vision processing algorithms, they are very challenging to employ for *real-time* vision processing, because of their high demand on computation and data movement [1]. It is well known that general-purpose processors cannot support CNN efficiently, because of their specific computational patterns [2]. Thus, various types of accelerators have been proposed based on Graphics Processing Units (GPU) [3], [4], Multiprocessor Systems-on-Chip (MPSoC) [5], [6], reconfigurable architectures [7]–[9], Field-Programmable Gate Arrays (FPGA) [2], [10], [11], analog circuits [12], in-memory computation [13], and dedicated hardware acceleration through Application Specific Integrated Circuits (ASIC) [14]–[17].

A typical CNN architecture consists of a stack of convolutional and pooling layers, followed by classifier layers, as shown in Figure 1(a). To realize *real-time* vision processing, all layers of the CNN should run on an accelerator. Otherwise, the data transfer time between the host and the accelerator cancels out the acceleration in the computation itself. The challenge is in the processing of the classifier layer, where all neurons are fully connected. Award-winning high-accuracy CNNs (such as AlexNet [18], which won the 2012 ImageNet contest) usually require a huge number of weights (up to 100s of MB [13]) and weights are not reused. The weights should be stored in an external memory (e.g., DRAM), and the performance is bounded by the memory access time [13].

This challenge is being addressed by recent CNN architectures. Two representative examples are SqueezeNet [19] and GoogLeNet [20]. SqueezeNet offers comparable accuracy to AlexNet, but it uses 510 times fewer weights. GoogLeNet took the first place in the 2014 ILSVRC Classification contest. GoogLeNet employs narrow layers to minimize the number of weights, while offering high accuracy by using a large number of such narrow layers (more than 100). As shown in Figures 1(b) and (c), the SqueezeNet [19] and GoogLeNet [20] architectures are not as regular as the traditional CNN architecture of Figure 1(a). AlexNet [18] and VGG-16 [21] are often used to evaluate prior work. Nevertheless, if the goal is to achieve high-accuracy vision processing, we believe SqueezeNet and GoogLeNet are good substitutes, because they offer comparable accuracy and are better suited to hardware acceleration due to their use of fewer weights.

To realize real-time vision processing, all layers of the CNN should run on the accelerator seamlessly. For example, Eyeriss [14], [22] requires reconfiguration of the accelerator for each layer. It takes 0.1 ms to configure one layer. If there are 100 layers, it takes 10 ms only for reconfiguration. ShiDianNao [15] addresses this by using hierarchical finite state machines. However, it is not proven with large-scale CNNs, such as SqueezeNet and GoogLeNet. Approaches using GPUs and FPGAs can execute all layers of the CNN quickly, but they consume an order of magnitude more power than ASIC designs. DaDianNao [23] offers low latency for all



(a) Traditional (b) SqueezeNet [19] (c) GoogLeNet [20]

Figure 1. Three different types of CNN architectures. The left one represents the traditional (generic) approach, while the other two represent two existing state-of-the-art approaches.

the layers of large-scale CNNs, but it consumes as much power as an FPGA, which may not be suitable for power-efficient vision processing. In general, an FPGA-based design cannot simply be implemented in an ASIC to boost power efficiency, due to the fundamental differences in the underlying design principles. Since the FPGA is programmable, the design can typically be customized to suit a particular CNN. This customization is not feasible in an ASIC. To support advanced CNNs like SqueezeNet and GoogLeNet in ASIC for real-time vision processing, we need a flexible – yet power-efficient – design that does not require run-time reconfiguration.

The proposed accelerator aims to achieve this goal by employing ***data-driven scheduling*** and ***modular design***. These two key features constitute *the novel contributions of this work*, since they enable the handling of *advanced CNNs without the need for reconfiguration*. The operation and destination of a Processing Element (PE) is determined at run-time upon receipt of data. The data is accompanied by metadata indicating the meaning of the data. By interpreting the metadata, a PE determines its schedule at run-time, which makes it easier to handle irregular CNN architectures. To achieve scalability, a modular design concept is employed with no shared resources and global synchronization being assumed. Each PE can only access its own local memory, and communicates only with its neighbors. Modular design facilitates deep pipelining, which enables further latency improvements by increasing the clock frequency. The accelerator has been enhanced from its original design [1] by employing on-chip memory optimization techniques, such as a sliding window and prefetching. As a result, it is demonstrated by experiments that the proposed accelerator executes all layers of SqueezeNet and GoogLeNet in 14.30 and 27.12 million cycles with 64 processing elements. Assuming a 1 GHz clock speed, these latencies correspond to 14.30 ms and 27.12 ms, respectively, which is comparable to high-performance FPGA-based approaches (range of 1.06 ms to 262.9 ms [24], [25]). It is estimated that the proposed accelerator consumes 2.47 W and 2.51 W for SqueezeNet and GoogLeNet, respectively, which may be higher than power-efficient ASIC-based approaches (consuming 0.278 to 0.320 W [15], [22]), but it is significantly lower than FPGA-based approaches (that consume 25 to 58 W [10], [24], [25]) and DaDianNao [23] (that consumes 15.97 W).

The rest of this paper is organized as follows: Section II discusses related work. After presenting the functional requirements and the architecture of the proposed accelerator in Section III, the details of the employed data-driven scheduling are explained in Section IV. In Section V, other salient features of the accelerator are described. Section VI provides experimental results, and Section VII concludes the paper.

II. Background and Related Work

Research in neural networks has a long history. Over the last several years, various types of approaches for the acceleration of CNNs have been studied. The design proposed in this paper relies on a fully digital ASIC implementation, using existing standard CMOS technology. We chose this approach, because it is practical (especially as compared to in-memory computation [13] and 3-D memory [26]), and we can potentially integrate a large number of PEs in a power-efficient manner (compared to FPGA implementations [2], [10]), as also acknowledged by [27]. The proposed accelerator can work with approximation [4], [28], compression [29], [30], and it can exploit the presence of zero weights [31]–[33].

There is a trade-off between latency and power consumption among these accelerators. The GPU approach achieves 0.19 ms latency at 227 W [34], while FPGAs offer a range of 1.06 ms to 262.9 ms at 25 W to 58 W [10], [24], [25]. These values are measured under AlexNet [18] or VGG-16 [21]. On the contrary, dedicated hardware accelerators implemented in ASIC target power-efficient implementations of small-scale CNNs, or the convolutional layers of large-scale CNNs [15], [27], [35], [36]. For example, Eyeriss [14] executes the convolutional layers of AlexNet [18] in 115.3 ms at 0.278 W [22].

Compared to two state-of-the-art CNN accelerators, the proposed accelerator offers lower latency and better scalability with the number of processing elements and clock frequency.

Compared to Eyeriss [14], the proposed accelerator offers significantly lower latency through its modular design (that allows for higher clock frequencies), weight prefetching (optimized memory access patterns to DRAM), and by using larger on-chip memory. Additionally, the data-driven scheduling enables seamless execution of all layers without reconfiguration. ShiDianNao [15] also supports seamless execution of all layers, by storing all weights and feature maps in on-chip memory. However, the ShiDianNao [15] architecture was evaluated only with small-scale CNNs whose weights and feature map sizes fit into on-chip memory. Furthermore, both Eyeriss [14] and ShiDianNao [15] employ global shared memory, which renders their scalability questionable. In contrast, the modular design concept of the architecture proposed in this work enables high clock frequencies through pipelining. Even though the proposed accelerator requires more hardware and memory space to accommodate its data-driven scheduling and modular design, it is still significantly more power-efficient than FPGA-based approaches.

III. Overview of the Proposed Accelerator

A. Functional Requirements

The current implementation of the proposed accelerator supports three types of layers, and four types of layer connections. The four layers are: (1) convolutional layer, (2) max pooling layer, and (3) average pooling layer. The classifier layer can be implemented as a special case of the convolutional layer. SqeezeNet and GoogLeNet still use the classifier layer, even though it is not as big as those in traditional CNNs. The pseudo codes of the three layers are shown in Figure 2.

```
for(row=0; row<R; row++)
  for(col=0; col<C; col++)
   for(ofm=0; ofm<M; ofm++)
    for(ifm=0; ifm<N; ifm++)
     for(i=0; i<K; i++)
      for(j=0; j<K; j++) {
       y = S*row+i;
       x = S*col+j;
       feature_map[layer][ofm][row][col] +=
        weights[ofm][ifm][i][j] *
        feature_map[prev_layer][ifm][y][x];
      }
```

(a) Convolutional layer

```
for(row=0; row<R; row++)
  for(col=0; col<C; col++)
   for(ofm=0; ofm<M; ofm++)
    for(i=0; i<K; i++)
     for(j=0; j<K; j++) {
      y = S*row+i;
      x = S*col+j;
      if(feature_map[layer][ofm][row][col] <
       feature_map[prev_layer][ofm][y][x])
        feature_map[layer][ofm][row][col] =
        feature_map[prev_layer][ofm][y][x];
     }
```

(b) Max pooling layer

```
for(row=0; row<R; row++)
  for(col=0; col<C; col++)
   for(ofm=0; ofm<M; ofm++) {
    for(i=0; i<K; i++)
     for(j=0; j<K; j++) {
      y = S*row+i;
      x = S*col+j;
      feature_map[layer][ofm][row][col] +=
       feature_map[prev_layer][ofm][y][x];
     }
    feature_map[layer][ofm][row][col] /= (K*K);
   }
```

(c) Average pooling layer

Figure 2. Pseudo codes of the 3 layers supported by the proposed accelerator. [R: Number of rows of the output feature map; C: Number of columns of the output feature map; M: Number of output feature maps; N: Number of input feature maps; K: Filter size; S: Stride. All of the R, C, M, N, K, and S are of the current layer.]

To support a traditional/generic CNN, only one type of layer connection is enough, which is shown in Figure 3(a). To support more advanced CNN architectures, the proposed accelerator supports three other types of connections. The feature maps of a layer can be split and sent to different layers, as shown in Figure 3(b), and all feature maps can be sent to multiple layers, as shown in Figure 3(c). Finally, output feature maps of different layers can be concatenated as input feature maps of a layer, as shown in Figure 3(d).

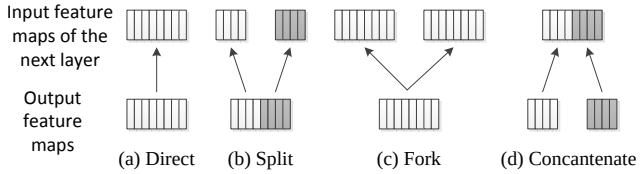


Figure 3. The 4 different types of layer connections supported by the proposed CNN accelerator that can be used to implement various CNN architectures.

The data-driven scheduling and modular design make it easy to support various types of layers and connections. Since the abovementioned three layers and four connections are enough to support SqueezeNet and GoogLeNet, the proposed accelerator only implements these for now, but it can be easily extended to cover other types of layers and connections. It is also possible to use heterogeneous PEs. These extension possibilities – and more – of the accelerator will be explored in our future work.

B. Overall Architecture

For real-time vision processing, the speed of the feed-forward process is more important than that of the backward process, because the backward process is usually performed off-line during training. Thus, the proposed accelerator is focused on accelerating the feed-forward process.

Figure 4 illustrates the architecture of the proposed accelerator and presents the high-level details of one PE module. We assume that the accelerator is implemented as a separate chip. It receives inputs from and sends outputs to the host through a standard bus interface. It has its own main memory (e.g., DRAM), which is used to store weights.

The proposed accelerator consists of a number of PEs. All PEs are the same, but one of them is designated as an interface PE, which interacts with the host and memory. The PEs are connected by 1D rings. Two rings are used for data (activation) transfer, and the third ring is used for weight prefetching. The details of the communication architecture will be explained in Section V-B.

A PE consists of a communication interface, matching logic, functional units (multiplier and adder), an output Finite State Machine (FSM), and local memories for weights and feature maps. The matching logic determines whether the incoming activation is assigned to the PE or not. The matching logic makes a decision based on the mapping information, which is presented in the next section (Section IV-A). If the

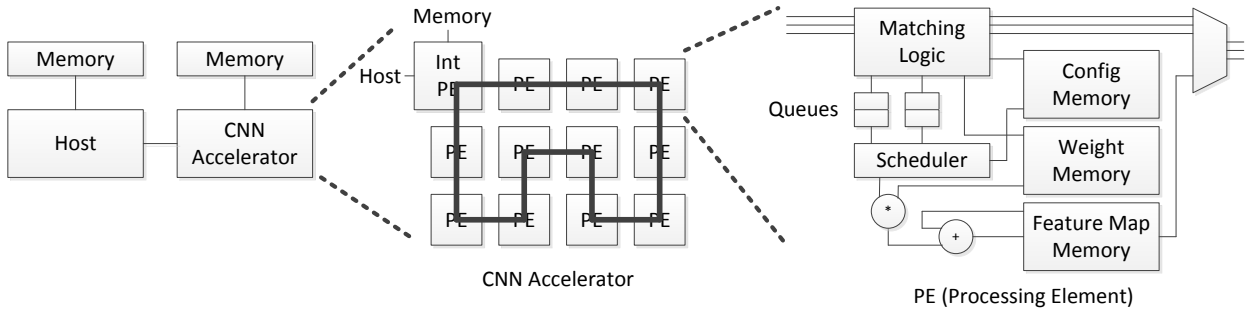


Figure 4. The architecture of the proposed accelerator and a high-level overview of one processing element. The pseudo codes of the 'Matching Logic' and the 'Scheduler' modules are presented, respectively, in Figure 6 and Figure 8.

Figure 5. Examples of message formats, including the pertinent metadata. [ec: Escape channel; ifm: Input feature map number; ofm: Output feature map number.]

incoming activation is accepted, it is pushed to a queue and processed by the functional unit. If the queue is full, the incoming activation cannot be accepted, even though it is destined to this PE. By interpreting the metadata accompanied by the activation, the corresponding functional unit is triggered. The result is stored in the local feature map memory, and transferred to other PEs when the computation is done.

IV. Data-driven Scheduling

The heart of the proposed accelerator and its key novelty is *data-driven scheduling*. It enables the execution of advanced CNN architectures without reconfiguration. Each PE determines whether to accept an activation and the subsequent schedule of operations, based on metadata and the CNN's configuration. The metadata is accompanied by the activation coming from the interconnection network. The CNN configuration is transferred from the host through the interface PE, and stored in the local configuration memory.

Figure 5 shows examples of the metadata. The format of the metadata depends on the type of data. For example, for activations, the metadata includes the layer, feature map, and the position (row and column) of the activation. To make the notation consistent with the pseudo code in Figure 2, the position of an activation in the input feature map is denoted as `y` and `x`, that of a neuron in the output feature map is denoted as `row` and `col`, and that of a weight in a filter is denoted as `i` and `j` throughout this proposal.

The configuration of layers is broadcasted to all PEs at initialization time, and it is stored in the local configuration memory of each PE. The configuration of one layer is shown in Table I.

The parameters R, C, M, N, K, and S are basic parameters of the CNN. Specifically, O and F are used to specify the connection, while F^{start} and F^{end} are used to support splits, and F^{shift} is used to support concatenation. For example, if a layer has 64 output feature maps, and 32 of them are sent

TABLE I. Configuration of a layer to be stored in configuration memory.

Parameter	Description
R	Number of rows of an output feature map
C	Number of columns of an output feature map
M	Number of output feature maps
N	Number of input feature maps
K	Filter size
S	Stride
O	Number of next layers connected with this layer
T_n	The layer number of $n-th$ connected layer
F_n^{start}	Start feature map number of the $n-th$ connected layer
F_n^{end}	End feature map number of the $n-th$ connected layer
F_n^{shift}	Feature map number shift of the $n-th$ connected layer

to layer 1, and the remaining 32 are sent to layer 2, then O=2, T_0=1, F_0^{start}=0, F_0^{end}=31, F_0^{shift}=0, T_1=2, F_1^{start}=32, F_1^{end}=63, and F_1^{shift}=-32. In this case, F_1^{shift} is used to convert the feature map numbers 32–63 of the current layer to the feature map numbers 0–31 of the next layer. In a similar way, when feature maps of multiple layers are concatenated, the feature map numbers can be adjusted to become linear, by using the F^{shift} parameter.

The rest of this section focuses on how data-driven scheduling is implemented.

A. Mapping

In the proposed accelerator architecture, the granularity of mapping is a feature map. A PE processes all neurons in its assigned feature maps. In this way, we can *avoid the sharing of weights among PEs*, which facilitates modular design. In other words, if a PE processes all the neurons of its assigned feature maps, it can store their weights in its local memory and other PEs do not need to access them. However, this mapping strategy may incur load imbalance, because it is inherently coarse-grained. The issue of load imbalance will be discussed in Section VI.

Feature maps are assigned as a combination of input and output feature maps. As a toy example, let us suppose a layer has 2 input feature maps (`ifm0` and `ifm1`), and 2 output feature maps (`ofm0` and `ofm1`). If there are 2 PEs, one PE is assigned to `ifm0-ofm0` and `ifm1-ofm0`, and the other PE is assigned to `ifm0-ofm1` and `ifm1-ofm1`. In other words, each PE processes all input feature maps of its assigned output feature map. If there are 4 PEs, feature maps are spread out as PE0 to `ifm0-ofm0`, PE1 to `ifm1-ofm0`, PE2 to `ifm0-ofm1`, and PE3 to `ifm1-ofm1`. PE0 and PE1 produce partial sums of neurons for `ofm0`, and one of them

```
ofm_start =
  index_start % M <= ifm ?
  index_start / M : index_start / M + 1;
ofm_end =
  ifm <= index_end % M ?
  index_end / M : index_end / M - 1;
if(ofm_end >= ofm_start)
  activation accepted;
```

Figure 6. The pseudo code of the matching logic. The code determines if an activation should be accepted or not.

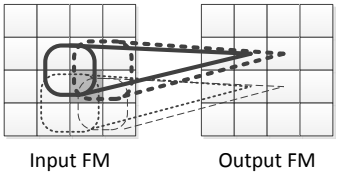


Figure 7. Illustration of how an activation is used for multiple filters.

must accumulate them. In the proposed accelerator, the PE processing the last input feature map of an output feature map is responsible to collect the partial sums from other PEs that are assigned to the same output feature map. In our toy example, PE0 should send its partial sums to PE1, so that PE1 can collect them and generate the final `ofm0`, while PE2 should send its partial sums to PE3, so that PE3 can generate the final `ofm1`.

To generalize this concept, we compute a feature map index for each combination of input and output feature maps, and a range of indices is assigned to PEs. The feature map index is computed as `index = ifm + ofm × M`, where `ifm` denotes the input feature map number, `ofm` is the output feature map number, and `M` is the total number of input feature maps. In the above toy example, the index of `ifm0-ofm0` is 0, `ifm1-ofm0` is 1, `ifm0-ofm1` is 2, and `ifm1-ofm1` is 3. If there are 2 PEs, PE0 is assigned to the range of indices from 0 to 1, and PE1 to indices from 2 to 3. If there are 3 PEs, PE0 is assigned to 0 and 1, PE1 to 2, and PE2 to 3. Thus, feature maps are not evenly distributed. If there are 4 PEs, each PE is assigned to each index.

The matching logic accepts an incoming activation, if its feature map falls within the range of the assigned indices. Recall that an activation is accompanied by metadata that includes the input feature map number, as shown in Figure 5. The pseudo code in Figure 6 shows how to determine if an activation, whose index is `ifm`, should be accepted or not, given a range of indices from `index_start` to `index_end`. Again, `M` indicates the total number of input feature maps.

Even if the activation is accepted, it should be forwarded to the next PE, because it may be used by the next PE. In fact, if there is a high enough number of output feature maps, as compared to the number of PEs, all PEs would need all input feature maps. Coming back to the toy example, let us suppose there are 2 PEs. PE0 processes `ifm0-ofm0` and `ifm1-ofm0`, while PE1 processes `ifm0-ofm1` and `ifm1-ofm1`. Thus, both PE0 and PE1 need all input feature maps (`ifm0` and `ifm1`). Therefore, we designed the accelerator in such a way that activations are broadcast, and PEs determine if they are to be accepted. This is in contrast to sending activations to specific target destinations.

Due to resource constraints, an activation may not be accepted, even if it is destined to the particular PE. Because of this, we need to maintain two types of counters. One counter is to determine when the activation should be removed from the network. When the activation is injected into the network, the total number of output feature maps is attached to the metadata. Whenever a PE accepts the activation, it decrements this counter by the number of assigned output feature maps and forwards it to the next PE. When this counter reaches zero, it is no longer forwarded (i.e., it is removed from the network).

The other type of counter is for determining if the activation has already been accepted, or not. Because a ring is used as a communication fabric in the proposed accelerator, the same activation may arrive at the PE more than once, if it is not removed from the network. To check for this, a PE maintains a counter for each input feature map of a layer. The activations of an input feature map are accepted in a pre-determined order. In our implementation, all columns of a row are accepted in an increasing order of their column index, and those of the next rows are accepted in the same way. The counter counts how many activations of the input feature map have been accepted. Since activations are accepted in a specific order, if a PE knows how many have been accepted, the PE can determine what should come next. The activation is accepted only if the incoming activation is what the PE is expecting. In this way, the PE avoids accepting the same activation more than once.

In case of the max and average pooling layers, the number of input and output feature maps is always the same. An output feature map only needs one corresponding input feature map. Thus, those PEs that generate the final output feature map of the previous layer (which is the input feature map of the pooling layer) are assigned to process the corresponding output feature map of the pooling layer. In this way, we can eliminate unnecessary activation transfers.

B. Scheduling

Once an activation is accepted, all operations that need the activation are scheduled. To compute a neuron, its neighboring activations are required. The exact number of required activations depends on the size of a filter. In other words, an activation should be used by multiple filters.

Figure 7 shows an example. Let us suppose the filter size is 2 by 2 and the stride is 1. To compute a neuron at [1][1] of an output feature map, we need activations (neurons of input feature map) at [1][1], [1][2], [2][1], and [2][2]. Similarly, neurons at [1][2], [2][1], and [2][2] of the output feature map need the same activation at [2][2] of the input feature map. If multiple output feature maps are assigned to the PE, neurons in other feature maps also need the incoming activation.

The pseudo code in Figure 8 shows how Multiply-And-Accumulate (MAC) operations are scheduled for an incoming activation. The `ofm_start` and `ofm_end` parameters are computed as shown in Figure 6. As shown in Figure 5, the position of the activation is given by `y` and `x`. The same mechanism is used for pooling layers. Instead of MAC operations, comparison (max pooling) or accumulation (average pooling) operations are scheduled.

The pseudo code is implemented as an FSM in the functional units. The FSM pops an activation from the queue located in-between the functional units and the matching logic in

```
for(ofm=ofm_start; ofm<=ofm_end; ofm++)
  for(row=MIN(y/S, R-1); row>(y-K)/S && row>=0; row--)
   for(col=MIN(x/S, C-1); col>(x-K)/S && col>=0; col--) {
     i = y-row*S;
     j = x-col*S;
     feature_map[layer][ofm][row][col] +=
      weights[ofm][ifm][i][j] *
      activation
   }
```

Figure 8. The schedule of operations when an activation is accepted. [R: Number of rows of the output feature map; C: Number of columns of the output feature map; K: Filter size; S: Stride. All of the R, C, K, and S are of the current layer.]

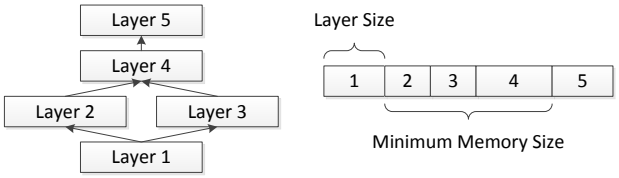


Figure 9. The minimum memory size requirement if parallel layers exist.

```
max = 0;
for(l=0; l<L; l++) {
   total_size=0;
   for all destination layers of layer l
     last = find the last destination layer;
   for(n=l; n<=last; n++)
     total_size += layer size of layer n;
   if(max < total_size)
     max = total_size;
   }
return max;
```

Figure 10. Pseudo code for determining the minimum memory size for each PE. The same pseudo code is used to determine the minimum size of the weight memory and the feature map memory. [L: Total number of layers.]

Figure 4. Once the FSM finishes all the scheduled operations, it pops the next activation from the queue. A functional unit accesses the weight memory and the feature map memory to perform its operation, and the result is stored in the feature map memory. To determine if accumulation is finished for one neuron, a counter is maintained for every neuron in the output feature map. The counter is stored in the feature map memory. The overhead of the memory will be discussed in Section VI.

V. Accelerator Design

This section presents the salient features of the proposed CNN accelerator that support data-driven scheduling.

A. Memory Optimization

Since on-chip memory is usually much smaller compared to the size of the weights and feature maps, a sliding window technique is adopted to manage on-chip memories. This technique is used for both the weight and feature map memories. Once all computations in a layer are complete, the window of memory slides so that the completed layer is freed from the memory, and the next layer is allocated. The PEs are not synchronized for the sliding process. Each PE decides to slide the memory independently from others.

In the case of the weight memory, a prefetching technique is used. All weights are stored in the external memory and they are prefetched while computation is progressing. Weights are stored in the order of layers and are always accessed sequentially. Thus, weights can be prefetched by taking full advantage of the throughput. Initially, the weight memory is allocated to the first couple of layers (the exact number depends on the weight size and memory size). As an example, let us suppose 2 layers are initially allocated. Before starting the computation, the weight memory is filled with weights of the allocated layers. Once the first layer completes, it is freed, and the third layer is allocated. While the PE is working on the second layer (using the weights of the second layer in the memory), the weights of the third layer are prefetched.

A memory needs to be large enough to store at least two layers: one for the source layer, and the other for the destination layer of the activation transfer. The destination layer of the activation transfer may not be the right next layer. As illustrated in Figure 9, there may be parallel layers in the CNN architecture. The proposed accelerator processes one layer at a time. In the example of Figure 9, the activations from layer 2 should be sent to layer 4. Thus, the weight and feature map memories need to retain memory space for layers 2 through 4. The algorithm to determine the minimum size of memory is shown in Figure 10.

The layer size depends on the type of memory. In case of the weight memory, the size of a layer is equal to the *filter size* $\times$ *number of assigned feature map indices* $\times$ *size of one weight*. Recall that the feature map index is a combination of input and output feature map numbers. The size of one weight depends on the number representation. We may use a 16-bit fixed-point number, or a 6-bit log value [28]. The proposed accelerator is not tied to any particular number representation. The feature map memory has two parts. One part is for output feature maps and the other is for counters. The layer size of output feature maps is equal to the *number of neurons of one output feature map* $\times$ *number of assigned output feature maps* $\times$ *size of one neuron*. The size of one neuron also depends on the number representation. Similarly, the layer size of the counters is equal to the *number of neurons of one output feature map* $\times$ *number of assigned output feature maps* $\times$ *size of one counter*. The size of one counter is log of *the filter size* $\times$ *the number of input feature maps*.

B. Communication Architecture

The proposed accelerator supports three types of communication patterns.

- **Broadcasting:** Since activations to convolutional layers are used by many PEs (often all PEs), they are broadcast. However, activations to pooling layers are not injected into the network, because they are processed by the same PE.
- **Single-source unicasting:** Weights are always sent from the interface PE. Since weights are not shared by PEs, one weight is sent to one PE (unicasting). Other initialization messages are also always sent from the interface PE.
- **Peer-to-peer unicasting:** The majority of traffic falls under the previous two types. If there are not enough output feature maps compared to the number of PEs, multiple PEs are assigned to the same output feature map, and partial sums need to be sent to a designated PE.

The traffic patterns are not required to preserve the message delivery order, i.e., the order of message arrival can be different

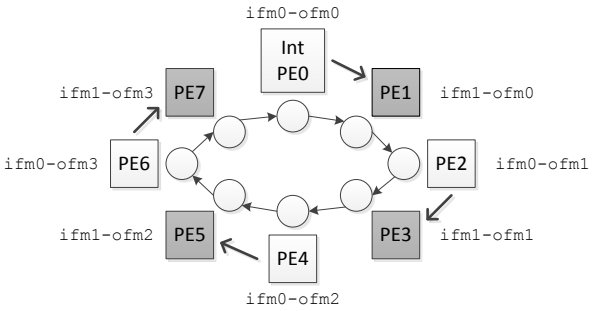


Figure 11. A logical view of the communication architecture of the proposed CNN accelerator. An example is given to illustrate how to minimize the hop distance under peer-to-peer unicasting.

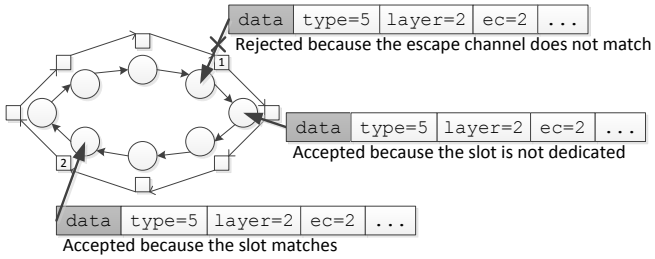


Figure 12. Employing *escape channels* ('ec') to avoid protocol-level deadlocks. Escape channels are implemented as slots.

from the order of departure for the same source-destination pair. Considering these patterns, we chose a uni-directional 1-D ring as a communication fabric, because it costs less than other packet-based Networks-on-Chip (NoC) [37]. The logical view of the communication architecture is shown in Figure 11. The topology is similar to that of Chain-NN [38], but we employ a ring instead of a systolic chain.

In the ring architecture, each PE interfaces with a ring stop. A message in the ring is ejected if it is destined for the local PE. Otherwise, it is forwarded to the next hop. A new message can be injected from the PE only if there is no message in the ring stop.

The ranges of feature map indices are assigned in such a way as to minimize the hop distance of peer-to-peer unicasting. As mentioned in Section IV-A, the PE processing the last input feature map of an output feature map is responsible to collect the partial sums from other PEs that are assigned to the same output feature map. If feature map indices are assigned in the same order as the topological order in the ring, the hop distance can be minimized. An example is given in Figure 11. In this example, `ofm0` is assigned to PE0 and PE1. Since PE1 processes the last input feature map, `ifm1`, it is designated to accumulate the partial sums and generate the final output. Since PE1 is located after PE0 in the ring topology, the partial sum sent from PE0 to PE1 takes only one hop.

To avoid protocol-level deadlocks, the concept of an *escape channel* is adopted. The ring itself does not cause network-level deadlocks, but because of the cyclic dependency caused by the upper-level protocol, deadlocks may occur. If the weight and feature map memories were infinite, there would be no chance of deadlocks, because the cyclic dependency would be broken by the memory. However, because the memory employs sliding, a later layer can only start when a previous layer completes, which forms a dependency from a later layer to a previous layer.

For example, let us suppose a CNN with 3 sequential layers and a PE have a feature map memory that is only enough to store two layers. At a certain moment, in PE0, layer 1 has completed, and activations are being transferred from layer 1 to layer 2. Layer 1 can be freed only after the transfer is complete, and layer 3 can then be allocated. In the proposed accelerator, PEs are not globally synchronized. Thus, another PE, say PE1, may have completed layer 2 and proceeded to layer 3. PE1 starts sending activations from layer 2 to layer 3. PE0 is also supposed to accept these activations for layer 3, but PE0 cannot, because the memory is not available. These activations can be removed from the ring only if all the assigned PEs accept them. Thus, until PE0 accepts these activations, they remain in the ring. The ring architecture employed by the proposed accelerator allows injection only if there is room in the ring. If the ring becomes full with these activations (from layer 2 to layer 3), PE0 may indefinitely not be able to receive activations from layer 1 to layer 2. This degenerate situation forms a cyclic dependency and causes deadlock.

To address this issue, an escape channel is introduced, and the escape channel is implemented as slots. As illustrated in Figure 12, slots are assigned and they are rotated in the ring. Each escape channel has one dedicated slot, and other slots can be used by all escape channels. Escape channels are assigned to layers. A different layer has a different escape channel. When an activation is generated, its escape channel is identified by the layer and put into the metadata, as shown in Figure 5. When an activation is to be injected, it can be injected only if its layer's escape channel matches with the slot. In this way, we can avoid the protocol-level deadlocks at a minimal cost.

The number of escape channels cannot be more than the number of ring stops (i.e., the number of PEs). In case of a deep CNN (e.g., GoogLeNet), the number of layers may be more than 100. A convolutional layer may need two escape channels, because it may need to send partial sums. Therefore, the number of escape channels should be optimized.

Since not all layers are active at a given time, we can reuse escape channels for the layers whose lifetime does not overlap. In fact, because of the data dependency between layers, no more than 3 layers can be processed at the same time. For example, let us suppose layer 1 is connected to layer 3 and layer 3 is connected to layer 5. Because there might be parallel layers, a layer may not be connected to the right next layer. The PEs are finished with layer 3 only if all activations are received from layer 1. Since the PEs are not globally synchronized, one of them may finish earlier than others and start layer 5. At this moment layers 1, 3, and 5 are active. However, because of data dependencies, no PE can process the layer after layer 5 until layer 3 completes. Since there are parallel layers between layers 1 and 5, the minimum number of escape channels is 10 in this case (2 of each layer times 5 layers), if all of the 5 layers are convolutional layers. Pooling layers do not need escape channels, because their activations are not injected into the network.

TABLE II. The default simulation parameters used in all experiments.

Parameter	SqueezeNet	GoogLeNet
Number of PEs	64	
Average memory access cycle	1	
Pipeline stages of communication channel	1	
Pipeline stages of functional units	1	
Queue depth	16	
Number of rings	3	
Configuration memory size	0.021 MB	0.092 MB
Weight memory size	1.289 MB	4.119 MB
Feature map memory size	9.132 MB	3.333 MB
Bit width of one activation ring	68	71
Bit width of the weight ring	58	61
Number of escape channels	10	46

TABLE III. Number of cycles required to execute all layers of the CNN.

CNN	Number of cycles	Execution time*
SqeezeNet [19]	14,303,612	14.30 ms
GoogLeNet [20]	27,122,439	27.12 ms

* 1 GHz clock frequency is assumed.

VI. Evaluation

A. Experimental Setup

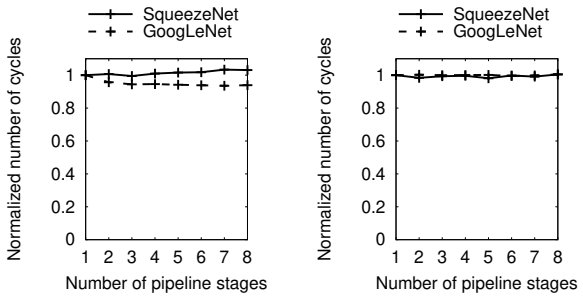
We developed a cycle-level in-house simulator using SystemC [39]. Since it is an architecture-level simulator, detailed analysis of the hardware cost is not available. However, we will discuss pipelining, which is related to clock speed, and the on-chip memory size, which has the most significant contribution to the hardware cost of the proposed accelerator. The default simulation parameters are shown in Table II.

The proposed accelerator can take full advantage of the DRAM bandwidth, because the access pattern is always sequential. All feature maps are stored in the on-chip memory by adopting a sliding window technique, and the external DRAM is used only for weights. Since weights are prefetched in the order of layers, there is no need for random accesses to DRAM. Assuming the proposed accelerator runs at 1 GHz, then a 2 GB/s throughput is required to fetch one weight (16 bits) per cycle. According to the DDR4 standard, the maximum throughput can be up to 25.6 GB/s. Therefore, the DRAM throughput is high enough to easily supply one weight every cycle.

B. Performance Analysis

Table III shows the number of cycles required to execute *all layers* of SqueezeNet and GoogLeNet. Under the assumption that the proposed accelerator runs at 1 GHz (since ShiDianNao [15] also runs at 1 GHz), these results correspond to 14.30 ms and 27.12 ms for SqueezeNet and GoogLeNet, respectively.

Even though a direct comparison may not be meaningful due to fundamental differences in the design goals (low power vs. low latency) and benchmark (different CNNs), Eyeriss [22] is reported to execute the convolutional layers of AlexNet in 115.3 ms, and the convolutional layers of VGG-16 in 4309.5 ms. While a GPU executes all layers of these CNNs in 0.19 ms, FPGAs require 1.06 ms to 262.9 ms [10], [24], [25], [34]. The performance of the proposed accelerator is comparable to FPGA-based techniques. DaDianNao [23] offers even lower latency, but its power consumption is comparable to FPGA-based techniques. This is because it targets high-performance implementations supporting all the layers of large-scale CNNs and both the forward and backward processing steps.



(a) Communication channel (b) Functional units

Figure 13. The number of pipeline stages does not have significant impact on the number of cycles required to complete the execution of the CNN. Thus, the performance of the proposed accelerator can potentially be enhanced by employing even higher clock frequencies.

It should also be noted that the proposed accelerator offers flexibility in that it can support SqueezeNet and GoogLeNet without run-time reconfiguration. Since SqueezeNet and GoogLeNet offer comparable accuracy with AlexNet and VGG-16, we believe they are good alternatives for power-efficient real-time vision processing.

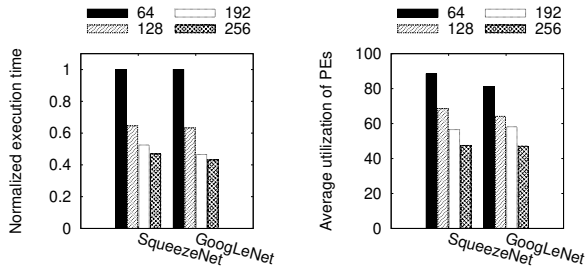
On the other hand, ShiDianNao [15] reports 0.047 ms to execute all layers of ConvNN [40]. However, ConvNN is much smaller. For example, GoogLeNet requires 1502 million MAC operations, whereas ConvNN only needs 0.6 million. While it demonstrates an efficient implementation of small-scale CNNs, it is not proven with large-scale CNNs for high-accuracy vision processing algorithms. Another previous work [35] is reported to execute a particular CNN in 20.55 ms, but said CNN is also small (20.81 million operations).

C. Scalability Analysis

If the budget allows, it is possible to further enhance the performance of the proposed accelerator by increasing the number of pipeline stages and the number of PEs. Figure 13 shows normalized number of cycles for SqueezeNet and GoogLeNet when the number of pipeline stages of the communication channel and the functional unit changes. In case of SqeezeNet, when the number of stages in the communication channel increases, there is a slight increase in the number of cycles. However, the increase is only 2.41% when the number of pipeline stages increases from 1 to 8.

In the case of pipelining the functional units, the pipeline may stall because of data hazards. However, all operations scheduled by an accepted activation are independent, because the operations are for different neurons. Data hazards happen only if there are overlapped neurons in the scheduled operations triggered by different activations. This probability is very low for convolutional layers. Even though the probability is relatively high for pooling layers, most of the cycles in the CNN are spent on convolutional layers. Therefore, the data hazards do not have significant impact on the number of cycles.

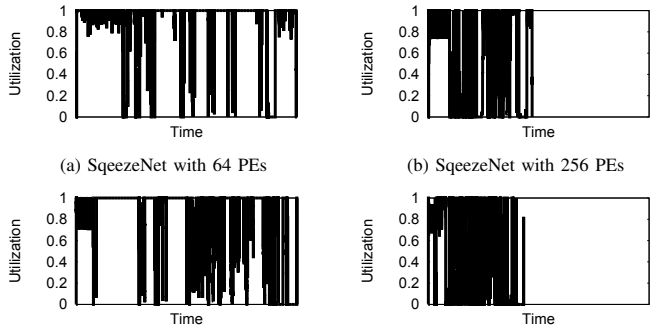
Figure 14 shows the normalized execution time and utilization rate when the number of PEs increases, up to 256. The execution time keeps decreasing and reaches 6.71 ms and 11.70 ms for SqueezeNet and GoogLeNet, respectively, when the number of PEs is 256. However, the utilization rate also decreases from 88.35% to 47.47% (SqueezeNet), and from



(a) Execution time varying the number of PEs

(b) Average utilization rate

Figure 14. The execution time keeps decreasing with increasing number of PEs, but the utilization drops gradually.



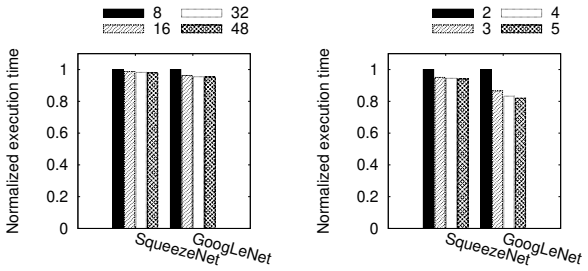
(a) SqeezeNet with 64 PEs

(b) SqeezeNet with 256 PEs

(c) GoogLeNet with 64 PEs

(d) GoogLeNet with 256 PEs

Figure 15. A utilization drop is observed in-between layers. This is attributed to load imbalance among the PEs.



(a) Queue depth

(b) Number of rings

Figure 16. Sensitivity analysis pertaining to the depth of the various queues and the number of employed interconnection rings. The chosen queue depth is 16 and the number of employed rings is 3, which are parameters shown to provide a good balance between performance and cost.

TABLE IV. The maximum supported values of the various CNN configuration parameters.

Parameter	Meaning	SqueezeNet	GoogLeNet
R	Rows	224	224
C	Columns	224	224
M	Input feature maps	1000	1000
N	Output feature maps	1000	1000
K	Filter size	7	7
S	Stride	2	2
O	Connections of a layer	2	4
T_n	Next layer	33	106
F_n^{start}	Start feature map	1000	1000
F_n^{end}	End feature map	1000	1000
F_n^{shift}	Feature map shift	1000	1000
Total number of layers		33	106
Total number of connections		40	204

81.03% to 47.10% (GoogLeNet). We found that this is not due to lack of scalability, but due to load imbalance.

Figure 15 shows the utilization rate over time for 64 and 256 PEs. As shown in this Figure, the utilization rate often hits maximum (100%) even when 256 PEs are used, which means the proposed accelerator is scalable in terms of the number of PEs. Comparing (a) versus (b), and (c) versus (d), we can observe a utilization drop, which is more frequent with 256 PEs than 64 PEs. The utilization drop is observed in-between layers. Though no global synchronization is assumed, the PEs cannot proceed to the next layer until other PEs finish their computation, because of data dependencies. Since our mapping strategy is coarse-grained, the workload may not be evenly distributed. If the number of PEs increases, the size of the assigned workload decreases, which makes the load imbalance relatively more significant. We will address this issue by fine-grained load-balancing in our future work.

D. Sensitivity Analysis

We determined the queue depth and the number of rings based on the sensitivity analysis shown in Figure 16. Specifically, our experiments indicate that a queue depth of 16 strikes a good balance between performance and cost. Similarly, 3 communication rings are seen as a cost-effective tradeoff. Recall that one of the rings is dedicated to prefetching weights.

E. Cost Analysis

To compute the *minimum* required memory size and the minimum required bit-width for the rings, it is essential to assess the *maximum* supported values of the parameters of the CNN configurations under investigation. These parameters are summarized in Table IV. The total number of layers used for the proposed accelerator is different from the number assumed in the original implementations of the CNN architectures. We slightly changed the architecture – in a mathematically equivalent manner – to better fit the underlying architecture of the accelerator. Specifically, instead of introducing an explicit concatenation layer, the output feature maps are directly connected to the next layer to reduce the memory requirement. Thus, if a pooling layer is followed by a concatenation layer, the pooling layer has to be split into the previous layers, because pooling layers are processed by the same PE where the output feature map is generated.

In the configuration memory, the basic parameters (R, C, M, N, K, S, and O) are stored for each layer and the connection parameters (T, F^{start}, F^{end}, and F^{shift}) are stored for each connection. The total number of bits to required to store all of these is 2,793 and 12,106 for SqueezeNet and GoogLeNet, respectively. Since all PEs need to store them, the sum of the configuration memory size of all PEs is 0.021 MB and 0.092 MB for SqueezeNet and GoogLeNet, respectively, as shown in Table II.

The minimum size of the weight and feature-map memories varies for different PEs, depending on the feature map assignment. For regularity, we used the same memory size across all PEs. The minimum memory size is computed as explained in Section V. The proposed accelerator does not depend on the type of number representation. All analysis results shown so far is based on 16-bit fixed-point representation, which is the most

TABLE V. The minimum required memory sizes under two different number representations.

Memory	SqueezeNet		GoogLeNet	
	16 bits	6 bits	16 bits	6 bits
Weight memory	1.289 MB	0.483 MB	4.119 MB	1.544 MB
Feature-map memory	9.132 MB	5.619 MB	3.333 MB	2.051 MB

popular setup in previous efforts. If, instead, we adopt 6-bit representation [28], the memory size can be further reduced. Table V shows both cases. Furthermore, since compression and pruning techniques [29]–[32] are also applicable to our accelerator, those techniques will be adopted in our future work.

Obviously, the memory size required for the proposed accelerator is significantly larger than that of existing accelerators. This is because the design goal of the proposed accelerator is to minimize latency as much as possible at a reasonable hardware cost. Considering the fact that recent Intel processors employ 8 MB of L3 cache and multiple 256 KB L2 and 32 KB L1 caches and DaDianNao [23] has a 36 MB embedded on-chip DRAM, we believe that 10 MB of on-chip memory is affordable for a stand-alone hardware-based CNN accelerator.

The longest message that the ring should carry is the activation, which is accompanied by the type of the message, the escape channel number, the layer number, the input feature map number, the position (`y` and `x`), and the counter, as shown in Figure 5. For future extensions, we assume 6 bits are used for the message type (i.e., 64 types of messages can be supported). The number of escape channels is computed as described in Section V, and is shown in Table II. The number of bits required to specify the layer, input feature map, and position can be calculated from Table IV. Since the maximum value of the counter is the number of output feature maps, 10 bits are assigned to this field. In total, 68 bits and 71 bits are required for one ring for SqueezeNet and GoogLeNet, respectively. For the ring employed for weight prefetching, the metadata includes the type of the message, escape channel number, layer number, input feature map number, output feature map number, and position (`i` and `j`), as shown in Figure 5. The total number of bits required for the weight ring is 58 and 61 for SqueezeNet and GoogLeNet, respectively.

F. Power Estimation

It is estimated that the power consumption of the proposed accelerator is similar to ShiDianNao [15], which consumes 320.10 mW (except for the memory power, which will be discussed shortly), assuming an operating frequency of 1 GHz. Both designs run at the same clock frequency, employ the same number of PEs (64), and use the same types of functional units (multipliers and adders). The overhead of the control logic would obviously be different, but according to the analysis in Eyeriss [22], the power consumption of the control logic corresponds to only 9.5% to 10.0% of the total power budget. In general, the biggest consumer of power is the on-chip memory. Since the proposed accelerator employs a significantly larger memory, it consumes more power than ShiDianNao, which has a 288 KB on-chip memory. By using the per-access energy model of CACTI [41] and the number of memory accesses obtained through simulation, the power consumption of both the on-chip memory and DRAM can be estimated. Including the power consumption of the other components reported by ShiDianNao, the total power consumption (including DRAM accesses) of the proposed accelerator is estimated as 2.47 W and 2.51 W for SqueezeNet and GoogLeNet, respectively. Despite the fact that these numbers are based solely on estimation, it is clear that the power consumption of the proposed accelerator is significantly lower than FPGA-based approaches (that consume 25 to 58 W) and DaDianNao's 15.97 W [23].

VII. Conclusions and Future Work

This paper proposes a novel hardware-based accelerator for deep CNNs used to realize *power-efficient real-time* vision processing. The new design achieves significantly lower execution latencies than existing power-efficient ASIC-based accelerators, primarily due to its inherent ability to operate at higher clock frequencies. This attribute is enabled by *modular design*, optimized memory access patterns due to *weight prefetching*, and larger on-chip memory. More importantly, the new accelerator can execute *all layers* of SqueezeNet and GoogLeNet in 14.30 ms and 27.12 ms, respectively, which are comparable to high-performance FPGA-based approaches, but with significantly lower power consumption at 2.47 W and 2.51 W, respectively. The use of *data-driven scheduling* can seamlessly support advanced CNN architectures without any reconfiguration. We expect that the proposed accelerator will expedite the widespread adoption of CNNs for power-efficient real-time vision processing, which is especially useful in the domains of unmanned vehicles, autonomous robotics, and surveillance cameras.

The data-driven scheduling scheme introduced in this work was applied only to CNNs. Nevertheless, we believe that it could also be used in other types of neural networks, and, specifically, in more recent networks, such as Recurrent Neural Networks (RNN) [42], Faster R-CNN [43], You Only Look Once (YOLO) [44], and Single Shot Detector (SSD) [45]. Moreover, if the load imbalance when the number of PEs grows beyond 256 is adequately addressed, we expect that the latency can be reduced even further.

References

[1] J. Lee and C. Nicopoulos, "A convolutional neural network accelerator for power-efficient real-time vision processing," in Proceedings of the The Twelfth International Conference on Advances in Circuits, Electronics and Micro-electronics (CENICS), 2019, pp. 25–30.

[2] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, "Optimizing fpga-based accelerator design for deep convolutional neural networks," in Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, ser. FPGA '15, 2015, pp. 161–170.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, June 2017, pp. 1137–1149.

[4] M. Imani, M. Masich, D. Peroni, P. Wang, and T. Rosing, "Canna: Neural network acceleration using configurable approximation on gpgpu," in 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), Jan. 2018, pp. 682–689.

[5] C. Wang, Y. Wang, Y. Han, L. Song, Z. Quan, J. Li, and X. Li, "Cnn-based object detection solutions for embedded heterogeneous multicore socs," in 2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC), Jan. 2017, pp. 105–110.

[6] F. Conti and L. Benini, "A ultra-low-energy convolution engine for fast brain-inspired vision in multicore clusters," in 2015 Design, Automation Test in Europe Conference Exhibition, March 2015, pp. 683–688.

[7] S. Cadambi, A. Majumdar, M. Becchi, S. Chakradhar, and H. P. Graf, "A programmable parallel accelerator for learning and classification," in 2010 19th International Conference on Parallel Architectures and Compilation Techniques (PACT), Sept. 2010, pp. 273–283.

[8] S. M. A. H. Jafri, T. N. Gia, S. Dytckov, M. Daneshtalab, A. Hemani, J. Plosila, and H. Tenhunen, "Neurocgra: A cgra with support for neural networks," in 2014 International Conference on High Performance Computing Simulation (HPCS), July 2014, pp. 506–511.

[9] M. Tanomoto, S. Takamaeda-Yamazaki, J. Yao, and Y. Nakashima, "A cgra-based approach for accelerating convolutional neural networks," in 2015 IEEE 9th International Symposium on Embedded Multicore/Many-core Systems-on-Chip, Sept. 2015, pp. 73–80.

[10] J. Qiu, J. Wang, S. Yao, K. Guo, B. Li, E. Zhou, J. Yu, T. Tang, N. Xu, S. Song, Y. Wang, and H. Yang, "Going deeper with embedded fpga platform for convolutional neural network," in Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, ser. FPGA '16, 2016, pp. 26–35.

[11] A. X. M. Chang and E. Culurciello, "Hardware accelerators for recurrent neural networks on fpga," in 2017 IEEE International Symposium on Circuits and Systems (ISCAS), May 2017, pp. 1–4.

[12] C. Mead and M. Ismail, Analog VLSI Implementation of Neural Systems. Springer, 2012.

[13] M. Gao, J. Pu, X. Yang, M. Horowitz, and C. Kozyrakis, "TETRIS: scalable and efficient neural network acceleration with 3d memory," in Proc. of the International Conference on Architectural Support for Programming Languages and Operating Systems, 2017, pp. 751–764.

[14] Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," in Proceedings of the 43rd International Symposium on Computer Architecture, ser. ISCA '16, 2016, pp. 367–379.

[15] Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, and O. Temam, "ShiDianNao: shifting vision processing closer to the sensor," in 2015 ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA), June 2015, pp. 92–104.

[16] M. Peemen, A. A. A. Setio, B. Mesman, and H. Corporaal, "Memory-centric accelerator design for convolutional neural networks," in 2013 IEEE 31st International Conference on Computer Design (ICCD), Oct. 2013, pp. 13–19.

[17] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," in Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems, ser. ASPLOS '14, 2014, pp. 269–284.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1097–1105.

[19] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size," CoRR, vol. abs/1602.07360, 2016.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014.

[22] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," IEEE Journal of Solid-State Circuits, vol. 52, no. 1, Jan. 2017, pp. 127–138.

[23] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, and O. Temam, "Dadiannao: A machine-learning supercomputer," in 2014 47th Annual IEEE/ACM International Symposium on Microarchitecture, Dec. 2014, pp. 609–622.

[24] X. Wei, C. H. Yu, P. Zhang, Y. Chen, Y. Wang, H. Hu, Y. Liang, and J. Cong, "Automated systolic array architecture synthesis for high throughput cnn inference on fpgas," in 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC), June 2017, pp. 1–6.

[25] U. Aydonat, S. O'Connell, D. Capalija, A. C. Ling, and G. R. Chiu, "An opencl™deep learning accelerator on arria 10," in Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, ser. FPGA '17, 2017, pp. 55–64.

[26] D. Kim, J. Kung, S. Chai, S. Yalamanchili, and S. Mukhopadhyay, "Neurocube: A programmable digital neuromorphic architecture with high-density 3d memory," in Proceedings of the 43rd International Symposium on Computer Architecture, 2016, pp. 380–392.

[27] C. Farabet, B. Martini, P. Akselrod, S. Talay, Y. LeCun, and E. Culurciello, "Hardware accelerated convolutional neural networks for synthetic vision systems," in Proceedings of 2010 IEEE International Symposium on Circuits and Systems, May 2010, pp. 257–260.

[28] D. Miyashita, E. H. Lee, and B. Murmann, "Convolutional neural networks using logarithmic data representation," CoRR, vol. abs/1603.01025, 2016.

[29] Y. Wang, H. Li, and X. Li, "Re-architecting the on-chip memory sub-system of machine-learning accelerator for embedded devices," in Proceedings of the 35th International Conference on Computer-Aided Design, ser. ICCAD '16, 2016, pp. 13:1–13:6.

[30] J. Zhu, Z. Qian, and C. Y. Tsui, "Bhnn: A memory-efficient accelerator for compressing deep neural networks with blocked hashing techniques," in 2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC), Jan. 2017, pp. 690–695.

[31] D. Kim, J. Ahn, and S. Yoo, "A novel zero weight/activation-aware hardware architecture of convolutional neural network," in Design, Automation Test in Europe Conference Exhibition (DATE), 2017, March 2017, pp. 1462–1467.

[32] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," in International Conference on Learning Representations, 2016.

[33] J. Albericio, P. Judd, T. Hetherington, T. Aamodt, N. E. Jerger, and A. Moshovos, "Cnvlutin: Ineffectual-neuron-free deep neural network computing," in 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), June 2016, pp. 1–13.

[34] nVIDIA, "Tesla m4 gpu accelerator," 2016.

[35] A. Dundar, J. Jin, B. Martini, and E. Culurciello, "Embedded streaming deep neural networks accelerator with applications," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 7, July 2017, pp. 1572–1583.

[36] A. Dundar, J. Jin, V. Gokhale, B. Martini, and E. Culurciello, "Memory access optimized routing scheme for deep networks on a mobile coprocessor," in 2014 IEEE High Performance Extreme Computing Conference (HPEC), Sept. 2014, pp. 1–6.

[37] J. Lee, C. Nicopoulos, H. G. LEE, and J. Kim, "TornadoNoC: a lightweight and scalable on-chip network architecture for the many-core era," ACM Trans. Archit. Code Optim., vol. 10, no. 4, 2013, pp. 56:1–56:30.

[38] S. Wang, D. Zhou, X. Han, and T. Yoshimura, "Chain-nn: An energy-efficient 1d chain architecture for accelerating deep convolutional neural networks," in Design, Automation Test in Europe Conference Exhibition (DATE), 2017, March 2017, pp. 1032–1037.

[39] SystemC, 2012. [Online]. Available: www.accellera.org

[40] M. Delakis and C. Garcia, "Text detection with convolutional neural networks," in International Conference on Computer Vision Theory and Applications, 2008, pp. 290–294.

[41] S. J. E. Wilton and N. P. Jouppi, "Cacti: an enhanced cache access and cycle time model," IEEE Journal of Solid-State Circuits, vol. 31, no. 5, May 1996, pp. 677–688.

[42] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.

[43] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," CoRR, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[44] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," CoRR, vol. abs/1506.02640, 2015. [Online]. Available: http://arxiv.org/abs/1506.02640

[45] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," CoRR, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

Novel Thermoelectric Energy Harvesting Circuit for Exploiting Small Variable Temperature Gradients Based on a Commercially Available Integrated Circuit

Martin Lenzhofer
Sensor Systems
Silicon Austria Labs GmbH
9524 Villach, Austria
martin.lenzhofer@silicon-austria.com

***Abstract*—Nowadays there is an increasing demand in generating electrical power out of ambient sources, such as light, wireless power, vibrations or heat, to supply low power electronics of self-sustaining sensor nodes. The main problem is that if dealing in harsh environments, or at places with very little energy budgets and additional varying environmental conditions, sophisticated harvesting circuits are necessary. In this work a novel approach for such a circuit is described that deals with thermal gradients. It bases on a commercially available integrated circuit, but with an external wiring scheme, according to the presented modified block diagram of a general harvesting system. A detailed description is given how to analyze the energy requirements. Additionally, the importance of adequate matching of the thermo-electric generator to the input impedance of the circuit is pointed out and at last the wiring of external components to override the internally fixed switching scheme of the build in blocks of the integrated circuit is described. Measurements verify the function of the low power thermal energy harvesting unit and demonstrate in detail the solutions to the key challenges to exploit maximal power out of thermal gradients and to supply a sensor node electronics over a certain time.**

Keywords-Harvesting circuit; thermal gradient; low power electronics; thermo-electric generator; thermal harvester; LTC3108.

I. INTRODUCTION

Recent developments in the field of wireless sensor technology and low power electronics enable an increase of the scope of applications in direction of self-sustaining sensor systems, even in harsh environments. Nowadays, most of the devices are generally powered by battery, but their drawbacks are that they increase the size of the devices and sometimes also their costs and pose an additional burden of replacement or recharging. Thus, there is an increasing effort in using other energy sources to supply the electronic circuits. The concept of energy harvesting describes the mechanism of exploiting energy from present sources in the environment of the sensor node. There are many ambient energy sources to exploit electrical power, but sophisticated circuits have to be developed to convert the small quantities of light emissions, wireless power transmitted by different devices, small vibrations or thermal energy into useful quantities that are sufficient to supply an electronic circuit [1]. A rough order of magnitude for the expected energy value for each energy source is given below.

Light energy is available almost everywhere and can be captured by photovoltaic (PV) cells, but it has to be considered that in average, levels between 10µW/cm² or just 10mW/cm² can be achieved indoor, depending on the location where the sensor node is used, meaning somewhere in the room or right beside a window, where the power levels are of course much higher, due to the fact of the sunlight influence. Of course, even higher values are achieved outdoor, if directly lightning the cell or photovoltaic module right adjusted to the sun, where values of up to 1.353W/m² [2] can be achieved, but these applications are not in the focus of the presented study dealing with energy harvesting solutions used in autonomous sensor nodes in a building or machine. To maximize the output power of the cell maximum power point tracking (MPPT) techniques must be applied. There are already commercial integrated circuits (IC) on the market with an implemented algorithm to continuously track the supply conditions and the corresponding load to maximize the transferred power, determined from the I-V curve of the solar cell.

Another omnipresent energy source is the radio frequency (RF) energy, due to the high number of mobile phones, WLAN networks, cell phone towers and every other kind of communication network. Beside the fact that the quantity of available power density level is very low, in the range of 0.1µW/cm² to 1µW/cm² [3], the energy levels additionally vary due to factors like terrain, number of users and many others. To achieve useful power levels also high gain and therefore big antenna systems are required and thus, often depict a limiting factor in being used in autonomous sensor nodes.

Also, the vibration energy source offers a great possibility to be converted into electrical power. There are many various forms - steady-state source, intermitted source and vibration source, where the last one is the commonly most often exploited one and being used to generate power, while for example human activity such as walking. While vibration amplitude can be quite large, like in the case of being exploited in civil structures ore railways, the expected levels, if being used under normal conditions, range between 4µW/cm² to 100µW/cm².

At last, thermal energy can be obtained either present in the ambient or generated during several processes. The most common way to generate electrical power to supply an electronic circuit of the available source is to use the Seebeck effect [2]. To extract energy from a thermal source a thermal gradient is required, where the conversion efficiency is directly related to the achieved temperature difference (hot and cold side) of the Peltier element or thermo-electrical generator (TEG). As it is described later in this paper, dependent on the application it is not always possible to achieve big gradients, leading in investigating novel design approaches however, to ensure the possibility of suppling an electronic circuit [3]-[5]. The expected energy output by exploiting thermal variations ranges between 10µW/cm² to about 60W/cm² [6]. As the temperature gradients are usually low, also the conversion efficiency and the output voltages of the systems are low, which afford special converter circuits. There are many applications, which focus on the usage of thermoelectric devices exploiting stable and continuous temperature differences, but there are rarely any of them dealing with harvesting energy from temperature gradients [7]-[10]. There are two main reasons for that: on one hand, the little energy potential within a gradient does not seem to be very attractive and, on the other hand, it is a reliability issue for stable operations of autonomous systems. Nevertheless, there are a plenty of applications that show small temperature gradients, for example, in wearable systems and in systems for environmental and infrastructure monitoring.

Referring to [1], the presented paper describes the functionality of the developed thermal harvesting circuit in more detail and additionally, demonstrates the possibility of exploiting energy from just a thermal gradient with a start-up procedure for an electronic circuit. Furthermore, the theoretical investigations point out the importance of matching the generator to the load impedance, as well as quantify the theoretical potential of a temperature gradient.

Section I describes the building blocks of a common energy harvesting system and the modified one. Section II deals with the development of the harvester circuit, based on the extended block diagram and Section III presents real measurements, verifying the functionality of this novel approach.

II. BUILDING BLOCKS OF AN ENERGY HARVESTING SYSTEM

Referring to the literature [11], commonly a conventional energy harvesting system is built up by several blocks, shown in Figure 1.

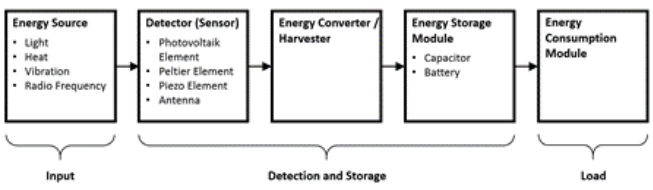


Figure 1. Block diagram of a conventional energy harvesting system.

At the input of the system the residual energy from the ambient environmental source is captured by the detector. There it is converted into the electrical domain and afterwards stored to supply different loads (circuit blocks). Optionally, as it was already mentioned in the introduction, some conditioning circuits can be developed to achieve a higher efficiency and output power. The storage module ensures continuous energy supply even under varying environmental conditions or also in the case, if the source is not available anymore.

This simple block diagram, shown in Figure 1, of an energy harvesting system is in general valid, but must be modified and extended, if dealing with not stable environmental conditions. Also, if working in the RF domain, where accumulation of energy is necessary, a more sophisticated system is needed. Referring to the described challenge of extracting energy from varying temperatures or its gradients, a similar approach to RF systems must be realized, to even have the ability of supplying any electronics. Figure 2 illustrates the modified block diagram for a system that is even able to address the previously mentioned challenges. Just a single reference could be found that describes this modified block diagram [12].

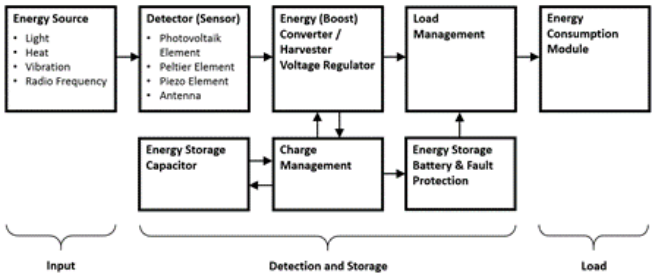


Figure 2. Modified block diagram for a general valid harvester system that is also able to deal with RF sources and thermal gradients.

Similar to Figure 1 the system includes a detector, represented by the TEG in this dedicated case, the voltage regulator working as energy converter, the charge management circuit to realize the energy storage block and additionally before connecting to the load the so important load management circuit. This last-mentioned block contains the key function for the developed harvester approach.

As most of the electronics require more than 1.8 V to operate adequately and often the input voltages are far smaller than that, a boost converter block is implemented. If connecting a TEG to the input, additionally variation in the temperature of the heat source can lead to an unstable output, which requires a regulation. The charge management block handles the different types of energy storage devices, which can either be an electrochemical cell (battery), or any kind of capacitor. Batteries offer a high energy storage density and less leakage current compared to capacitors, but therefore cannot provide high current bursts for a short period of time.

The load management circuit is used to automatically switch the load between the regulated voltage and the storage device, depending on the environmental conditions and the status of the regulator output. In case of using a battery, also a protection block to prevent overcharging or over discharging is implemented in the block diagram.

III. Development of the Harvester Circuit

If dealing with very little energy, as it is contained in thermal gradients, one of the most important points is an adequate impedance matching for maximal power transfer. Additionally, an energy storage system must be considered, to accumulate enough energy before powering-up the connected sensor node electronics based on an intelligent start-up sequence switching scheme.

A. *Impedance Matching for Maximal Power Transfer*

Thermal energy harvesting systems require specific optimizations, either on the thermal level and on the other hand on the electrical circuit level. In particular, the thermoelectric device must be sized appropriately for the available heat and the electrical load. The presented work bases on the already optimized thermal setup and therefore just focus on the electrical domain of building up a harvesting circuit to supply a sensor node electronic circuit. Through the thermoelectric coupling equations, the open circuit output voltage of the TEG, described as U_{TEG} is given by [13],

$$U_{TEG} = \alpha \cdot \Delta T_{TEG} \tag{1}$$

with the Seebeck coefficient α and the heat difference ΔT_{TEG} between the hot and the cold side of the TEG, including of course the contact and thermal interface resistances of the build-up mechanical system. The unloaded TEG output voltage U_{TEG} varies linearly with the temperature difference. Figure 3 shows the measurement result U_{TEG} (without load) over time of the TEG system during the heating up process, when the electronic circuit should work, and the converter circuit is designed for.

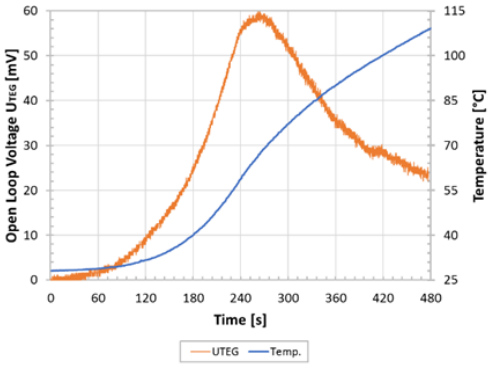


Figure 3. Open loop voltage U_{TEG} of the TEG versus the temperature raise in the oven over time. After about 260s the thermal mass of the sensor node is heated up, leading to a further decrease in voltage again.

The resulting current flow through a fictive resistive load R_L can be described as [14],

$$I_L = \frac{\alpha}{R_L} \cdot \frac{\Delta T_{TEG}}{1+\frac{1}{u}} \tag{2}$$

with,

$$u = \frac{R_L}{R_{TEG}} \tag{3}$$

where R_{TEG} is the electrical resistance of the TEG and u the ratio between the TEG impedance and the load impedance R_L. If applying formula (1) and (3), the resulting current I_L, expressed by (2), through the circuit becomes as,

$$I_L = \frac{U_{TEG}}{R_L+R_{TEG}} \tag{4}$$

As the current I_L flows through the internal impedance R_{TEG} and through R_L, the maximal voltage drop is achieved if both values are of the same value, or the ratio u = 1. In this case the maximum amount of energy can be transferred, and appropriate power matching is performed, which represents the aimed condition.

In this development, the situation is more complicated, because several aspects must be considered. First, Figure 3 just illustrates the voltage of the TEG generated during a heat up ramp without any load. Before applying this result to (4) also the real impedance of the TEG is measured with an LCR bridge, at a frequency of 100 kHz, which value is equal to the switching frequency of the connected DC-DC converter circuit. It turns out that the internal resistance of the TEG is about 5.4 Ω and additionally there is a small inductance, caused by the wires and the internal cabling of about 2 µH. Furthermore, also the input impedance of the DC-DC converter unit is analyzed with a source meter according to the input voltage level, shown in Figure 4. The discontinuity in the curve at about 25 mV represents the level of start-up of the circuit. From the harvesting point of view this means that for about 230 s the voltage just increases to that level, which represents the moment of reaching twice the value of the unloaded result in Figure 3 performed at the state of perfect power matching. After that, energy is transferred to a storage element, which is just possible for a certain time, in this particular case just as long as the system does not reach +85°C, which occurs after approximately 330 s. Higher temperatures are not allowed, due to the fact of proper reliable functionality of the circuit. Figure 5 illustrates the temperature raise of the oven versus the achieved temperature difference over the TEG, which represents a linear correlation of the generated voltage U_{TEG} in Figure 3. Due to the fact of still increasing temperatures in the oven, the whole sustainable sensor setup will heat up, resulting in a decreasing voltage difference over the implemented TEG. So, there is just a limited time-window to supply the electronics and perform the processing ability of the sensor

node. In the timeframe from the start-up of the converter to the maximal allowable temperature, enough energy to perform all the processing for the controller unit must be collected. Also, all the losses and the conversion efficiency of the converter unit must be considered.

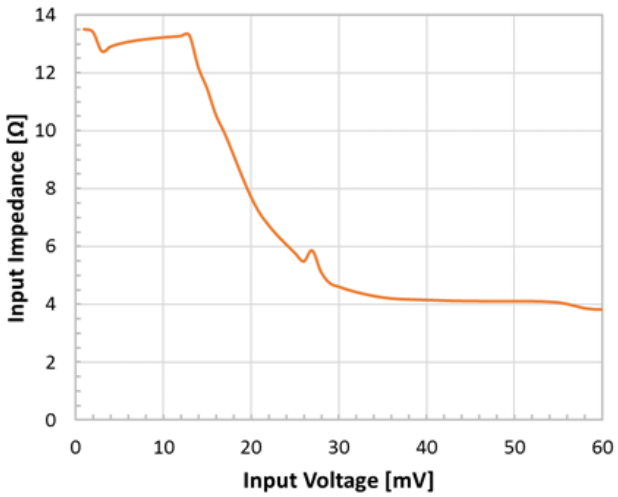


Figure 4. Measurement of the input impedance of the DC-DC converter circuit, dependent on the voltage level on its input. The discontinuity in the impedance chart around 25 mV represents the start-up condition of the electronic circuit.

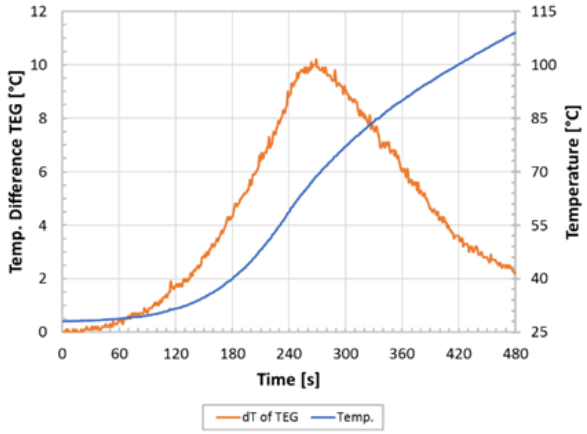


Figure 5. Temperature raise of the oven (blue curve) and temperature difference between the hot and the cold side of the TEG in the setup (orange curve).

Putting all these dependencies together and additionally applying formula (4) the maximal available electrical power at the input of the DC-DC converter can be expressed by,

$$P_L\left(t, U_{TEG}(t)\right) = \left(\frac{U_{TEG}(t)}{R_L\left(U_{TEG}(t)\right)+R_{TEG}}\right)^2 \cdot R_L\left(U_{TEG}(t)\right) \quad (5)$$

Figure 6 shows the result of equation (5) (blue signal) and additionally the respective integral of the signal (orange), which leads to the available energy value at the circuit input to supply the sensor node electronic circuit. Still it must be considered that this power is available at the input, and additional conversion losses of the circuit itself must be added and determined.

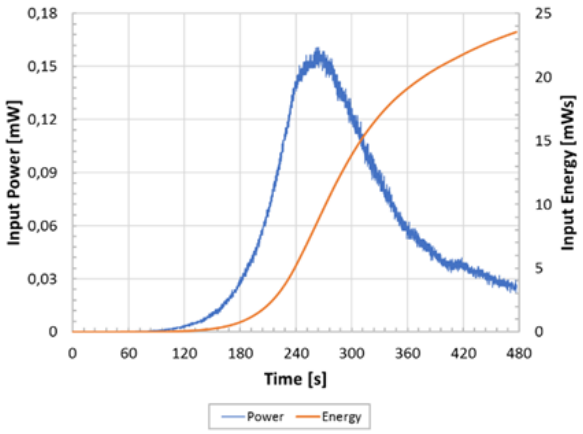


Figure 6. Blue signal represents the available input power and the orange signal represents the integral and therefore the energy over time at the circuit input.

There are just a few vendors of integrated ultra-low power converter and management IC on the market, where the LTC3108 [14] from Linear Technology Inc. seems to be the best fitting one for the present use case. This circuit is intended to provide a highly integrated DC-DC converter for harvesting and managing surplus energy from extremely low input voltage sources, such as the described TEG system. Referring to the datasheet [14], the implemented step-up technology operates from input voltages as low as 20 mV. This minimum input voltage depends on the transformer turns ratio, the load power required and the internal DC resistance (ESR) of the supplying source. A lower ESR allows the use of lower voltages and provides higher output power capability, but in the described use-case, the limitation is given by the TEG's resistance of 5.4 Ω resulting in a circuit start-up voltage level around 25 mV, shown in Figure 4.

B. Wiring scheme of the DC-DC Converter according to the modified block diagram

The recommended wiring scheme for a wireless remote application is given in the datasheet [15], shown in Figure 7. Since this is just working for continuous applied input voltage levels, some modifications must be applied, to realize the suggested wiring scheme according to the extended block diagram shown in Figure 2.

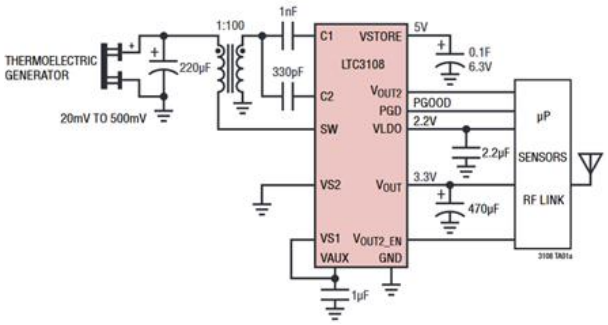


Figure 7. Wireless remote sensor application powered from a Peltier cell, according to the datasheet suggestion. This wiring scheme represents a solution according to the typical block diagram, shown in Figure 1.

Before thinking of how to change and modify the connection scheme of all the implemented blocks within the LTC3108 it is necessary to have a deeper look into the sequencing of the chip, refer to Figure 8, considering that this is always, just valid for stable and constant input conditions.

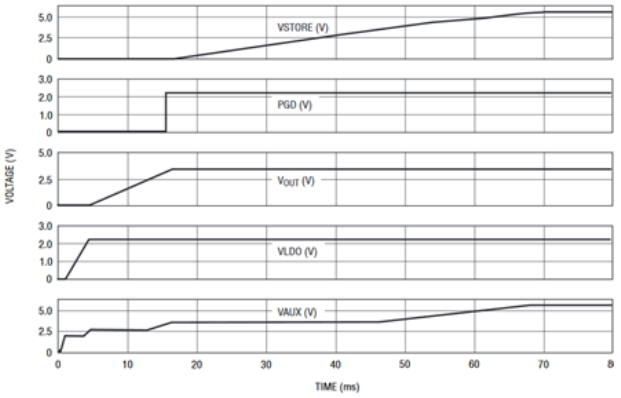


Figure 8. Start-up sequencing diagram of the management unit of the LTC3108 from Linear Technology Inc..

If a voltage is applied to the input of the circuit the level at VAUX starts to increase. The precision internal micropower voltage reference to accurately control the output voltage level V_{out} becomes active, after exceeding a level of 2 V. At this point, the synchronous rectifiers, which are switched in parallel to each diode, take over leading to an improvement in the conversion efficiency. After reaching a VAUX level of 2 V the VLDO supply starts-up, which provides 2.2 V output for powering low power processors or other low power ICs, if dealing with constant input conditions. Afterwards the level at Vout starts to increase till the output level is reached, which is programmed via the configuration pins VS1 and VS2 and set to 2.35 V in this dedicated case. If the level is valid the power good PGD indicator pin is activated and the internal charge control unit switches to another output to store the energy into an external capacitor connected to the pin VSTORE. The level will still increase up to 5.5 V and will remain at this level during the time where the V_{in} level is present. The whole routine will be processed within a few milliseconds at stable input conditions.

If using the implemented blocks of the semiconductor device to exploit temperature gradients, the switching scheme of the charge control circuit must be influenced from outside, to affect the hardware implemented start-up sequence. One major problem if working with thermal, time and slope varying gradients is the very limited contained energy. In the case of constant heat transfer of the TEG, load matching is targeted and enough to guarantee maximal power transfer. If dealing with that small temperature gradients this configuration is not enough, due to the fact that already a low-power load like a Microcontroller Unit (MCU) core connected to the Low-Dropout Regulator (LDO) output of the chip, the current flow leads to a decrease in voltage at the input. Tests showed that even for very low power MCUs, a valid start-up of the controller core is impossible. Therefore, another strategy must be implemented; meaning, first to harvest all available energy within the thermal gradient and just activate the controller circuit at the state of enough available power.

Figure 9 shows the adopted schematic scheme to deal with positive variable temperature gradients that result in a positive voltage level at the TEG system, which is suitable if using the sensor in an oven within the start-up procedure.

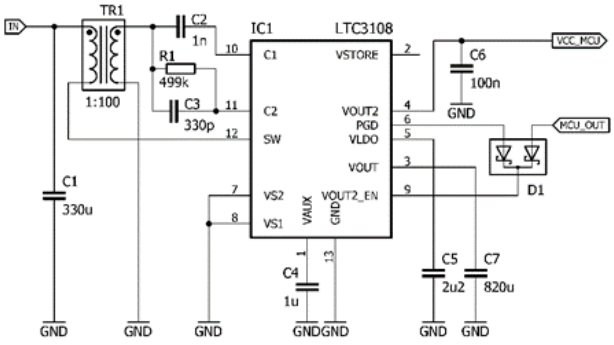


Figure 9. Developed circuit of the thermal harvester unit with improved power-up sequence influenced by external components.

If an input level of around 20 mV is reached, the DC-DC converter of the LTC3108 starts-up, leading to a raising voltage level at the pin VOUT. To avoid a constant current flow and perform adequate matching, just a capacitor as load is connected to this pin. This leads to a slowly charging of the energy storage device and capturing all available power from the input. The capacitance value must be matched to the slope of the thermal gradient, because a continuously heating up system, will result in a decrease of level over the TEG after a certain time.

If remaining heating-up of the whole self-sustaining sensor node, the DC-DC converter will be deactivated again, if turning back below the approximately 20 mV. In the case of a too large capacitance value, the expected output voltage level at VOUT will not be reached and therefore PGD is not set. The stored energy in this capacitor can be used to supply the sensor electronics afterwards. The IC internally generates a PGD signal, if the programmed output level at VOUT reaches 2.35 V. In the presented application, the load formed by an MCU is connected to the pin VOUT2. The reason is that this pin is switchable by an internal low leakage transistor. For enabling this VOUT2 pin, the PGD signal is now used and connected to the respective input pin with a low voltage Schottky diode D1.

This diode D1 represents another key aspect of the presented circuit, because it forms together with the second one an OR-gate, controlled by an output pin of the external MCU. This method is necessary due to the load case at output VOUT2. The start-up current of the MCU, refer to Figure 10 a.) (state A), leads to a voltage drop at VOUT (state B), where the energy flows from the capacitor C7 over the internal transistor of IC1 to the MCU. The PGD condition of the LTC3108 is tied to an internal not access- and adjustable hysteresis setting, which leads to switch off again, if the voltage level decreases by 7.5% of the nominal set value (state B). This happens quite rapidly, because the start-up current even of a low power MCU is quite high in the range of a few mA, or approximately 1.7 mA in the presented application. If the supplying voltage VOUT2 is switched off by the PGD signal again, the electronics would never reach the properly run mode of the circuit electronics. After switching off again (state B), the load current is almost zero again, causing an increase in voltage at VOUT again and switching on the voltage regulator once more (state C). So, this effect of switching on and off the load leads to a ringing phenomenon, refer to Figure 10 a.), which must be prohibited.

Therefore, the energy stored in the capacitor C7 must be high enough to hold the voltage level during the start-up of the MCU, configuration and controlling the diode D1 above the rated voltage (set voltage level reduced by 7.5% representing the internally fixed hysteresis value). After that, a further decrease in voltage level is not critical anymore, due to the locking function of the high level of the MCU pin that deactivates the hysteresis because of the OR gate mechanism, resulting to the diagram shown in Figure 10 b.).

In the presented application, the output level V_{out} is set to 2.35 V, so the start-up procedure including the activation of the output pin is not allowed to decrease beyond 2.174 V. As an MCU with a supply rating of 1.8 V is used, still a margin for a further approximately 0.4 V voltage drop is guaranteed for valid operation. Depending on the stored energy, the permitted maximal voltage drop and the current demand, which is of course mainly influenced by the software control algorithms and its necessary internal hardware blocks of the controller, the active operation time of the system, can be evaluated.

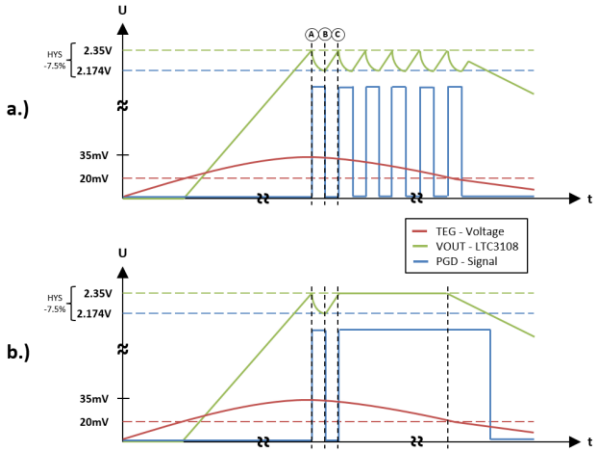


Figure 10. a.) Ringing phenomena due to different load conditions at VOUT, b.) Avoidance of the effect through self-locking circuitry via the MCU controlled OR gate configuration.

IV. Measurements and Validation

In this section, the developed circuit is evaluated. Compared to the measurement result of the open-loop voltage of the TEG, shown in Figure 3, it turns out that almost perfect maximal power matching is achieved, due to the fact of reaching nearly half of the voltage level. The measurement is shown in Figure 11 and it turns out that the harvesting part of the circuit starts-up at 21 mV, leading to an increase in the voltage level at VOUT. After the set output voltage level of 2.35 V is reached, the circuit tries to hold the level. Because the input voltage starts decreasing again the bursts in the signal at VIN becomes higher and the frequency decreases. At approximately 320 s the DC-DC converter stops working. Since there is still energy stored in the capacitor connected to VOUT, the level remains stable for a time. Referring to the measurement in Figure 3, at this time also a temperature of about +85°C is reached, meaning that the circuit will not work much longer. Because of the thermal mass of the sensor node, the inner temperature is slightly lower, leading the circuit to extend the working time.

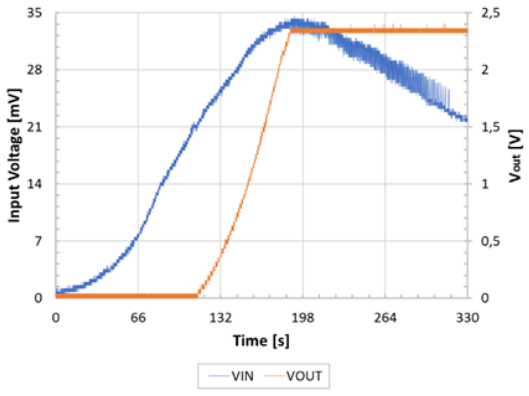


Figure 11. Blue signal: Input voltage level at the DC-DC converter input under load matching condition; Orange signal: Output voltage slope at pin VOUT achieved during heat-up process based on input level of the circuit.

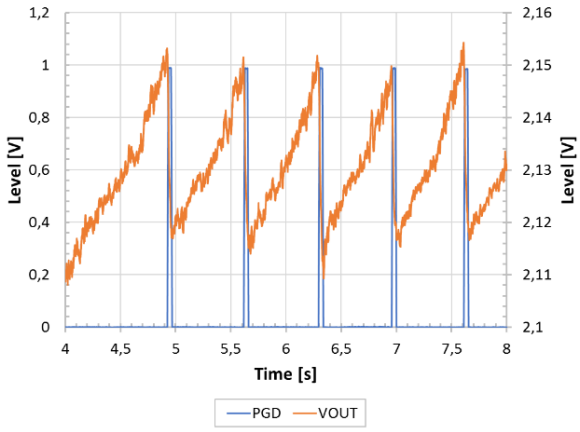


Figure 12. Circuit behavior without self-locking mechanism. If PGD (blue signal) is switched on, the circuit is supplied, leading to a voltage drop at VOUT (orange signal) caused by the current flow of the MCU and PGD is switched off again, if the value is lower than the internal hysteresis value. Due to the no-load condition after switching off, the capacitor charges up again and the voltage level increases.

Because the start-up current of the electronics is quite high, the voltage drops rapidly under the internal hysteresis value leading the LTC3805 to switch off again the PGD pin. Therefore, the circuitry does not have enough time to start-up properly. The active ON time is limited to the burst length, which is just 34 ms in this dedicated case. After switching off PGD, the current supply of the connected electronics stops, leading capacitor C7 at pin VOUT to charge up again. The desired and programmed output voltage is reached again after approximately 660 ms, refer to Figure 12. Now PGD is switched on again leading to repeat this sequence. Since the input voltage decreases, also the charge-up time of the capacitor is influenced, leading to longer charging times after a while. This kind of ringing phenomena makes it necessary to evaluate the function of the developed self-locking circuitry.

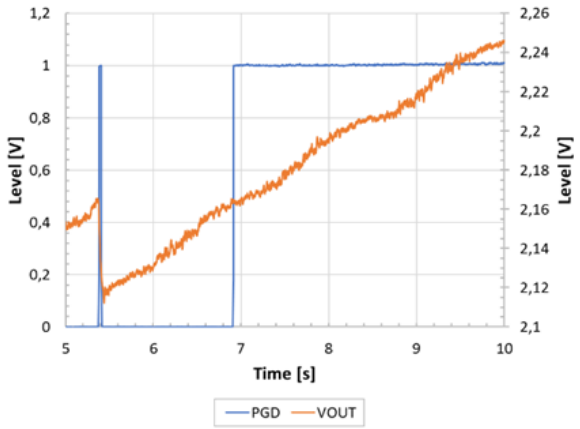


Figure 13. Working principle of the self-locking mechanism. Because the MCU switches on a pin to activate the OR gate leading to keep supplying the circuit, the PGD switches on after reaching the valid VOUT level again.

Comparing the time between the first and the second activation of PGD, shown in Figure 13, the time duration is longer than the one measured without the self-locking mechanism in Figure 12, which is approximately 660 ms long. The capacitor C7 needs approximately 1.5 s to charge up at the given input conditions and to reach the valid output level. This is the reason for implementing the self-locking mechanism, because the MCU already activates the pin to switch the OR gate configuration right after initialization, which is manageable to be done within a few milli seconds. The pulse time of the PGD signal is about 34 ms long, as pointed out in Figure 12. Within this time the MCU must take over the function of holding the level at the enable pin that switches on the supply of the electronics at VOUT2. Nevertheless, there is a trade off in dimensioning the capacitor C7 connected to VOUT. On one hand, it must provide enough energy to start-up the MCU and, on the other hand, it must not be too large, because then a charging up to the desired level, of 2.35 V in this case, is not possible anymore. So, the value of capacitance must be optimized according to the power consumption of the MCU or sensor node electronics and the input voltage VIN provided by the TEG matched to the input impedance of the DC-DC converter circuit. Additionally, the internal series resistance of the capacitor C7 should be as low as possible to provide high current pulses when switching on the MCU.

To sum up, if designing a harvesting unit for an electronic circuit that must deal with variable input conditions as it is given in the RF domain or even if dealing with thermal gradients, it is necessary, first to analyze the electrical energy needed, to measure the current afford necessary to start up the circuit and to investigate into the available energy source. Next, it must be tried to dimension the detector element, the TEG in this presented use case, according to the desired power budget and to perform best possible input matching. In a next step, the introduced modified block diagram must be applied to develop the circuit and finally the energy storage element must be matched, in respect to capacitance and technology, as described before. If all these stated items are considered it is possible to develop a harvesting circuit that is even possible to supply an electronic circuit for a certain time span under varying environmental conditions and in situations when very little energy is available, even under harsh environments.

V. CONCLUSION

This paper briefly introduces an overview of possible power sources, specially focusing on thermal sources that can be used to generate energy for a self-sustaining autonomous sensor node. Specially in harsh environments and closed compartments, like ovens, battery-based systems cannot be used. Therefore, thermal harvesting units that exploit the energy from the heating-up process in the oven are necessary. Since the contained energy potential is very limited, as it turns out of the theoretical analytics, as well as compared to the carried-out measurements, the presented

novel approach deals with an optimized circuitry based on a commercially available harvesting IC. Externally applied components influence the implemented switching sequence and therefore realize a circuit based on the presented modified harvesting block diagram. Beside adequate power matching of the energy converter, or thermal electric generator in the presented use case, and the input impedance of the connected step-up converter, also an intelligent load management is necessary that guarantees in a first step to accumulate all the available energy and just after a certain time, to activate the processing unit or sensor node electronics. Without such switching mechanism, already a low power MCU that is instantaneously supplied, would cause a current flow that would not allow the electronics to start-up adequately, or even reach a proper and stable voltage level. Therefore, this paper presents a solution to solve these challenges, of harvesting little power from a thermal gradient, but also enabling a proper operation of a sensor node electronics over a certain time. Finally, measurement results prove the proper functionality of such a low power thermal energy harvesting unit electronics.

ACKNOWLEDGMENT

This work was performed within the COMET Centre ASSIC Austrian Smart Systems Integration Research Center, which is funded by BMK, BMDW, and the Austrian provinces of Carinthia and Styria, within the framework of COMET - Competence Centres for Excellent Technologies. The COMET programme is run by FFG.

REFERENCES

[1] M. Lenzhofer, “Thermoelectric Energy Harvesting Circuit for Small Variable Temperature Gradients”, IARIA Proceedings of SENSORDEVICES 2019, The 10th International Conference on Sensor Device Technologies and Applications, 27.-31.10.2019 Nice, France; ISBN: 978-1-61208-745-0, pp. 61-62, 2019.

[2] F. Yildiz and K. L. Coogler, "Low Power Energy Harvesting with a Thermoelectric Generator through an Air Conditioning Condenser", 121st ASEE Annual Conference & Exposition, Indianapolis, vol. 10552, 2014.

[3] D. Zabek and F. Morini, “Solid state generators and energy harvesters for waste heat recovery and thermal energy harvesting”, Thermal Science and Engineering Progress, vol. 9, pp. 235-247, ISSN 2451-9049, 2019, https://doi.org/10.1016/j.tsep.2018.11.011.

[4] A. M. Abdal-Kadhim and K. S. Leong, "Application of thermal energy harvesting from low-level heat sources in powering up WSN node”, 2nd International Conference on Frontiers of Sensors Technologies (ICFST), Shenzhen, 2017, pp. 131-135, doi: 10.1109/ICFST.2017.8210489.

[5] P. Mullen, J. Siviter, A. Montecucco and A. R. Knox, “A thermoelectric energy harvester with a cold start of 0.6 °C”, 12th European Conference on Thermoelectrics, Materials Today: Proc. 2, pp. 823–832, 2015.

[6] F. Yildiz, “Potential Ambient Energy-Harvesting Sources and Techniques”, Journal of technology studies, vol. 35, no. 1.

[7] V. Leonov, P. Fiorini, S. Sedky, T. Torfs and C. Van Hoof, “Thermoelectric MEMS generators as a power supply for a body area network”, Proc. IEEE Transducers 2005, pp. 291-295, 2005.

[8] Y. Meydbray, R. Singh, and A. Shakouri, “Thermoelectric module construction for low temperature gradient power generation”, Proc. 24th Int. Conference on Thermolectrics, pp. 348-351, 2005.

[9] P. Woias, “Thermoelectric Energy Harvesting from small variable Temperature Gradients”, Proc. 12., Dresdener Sensor-Symosium, pp. 83-88, DOI 10.5162/12dss2015/5.6, 2015.

[10] A. Moser, M. Erd, M. Kostic and K. Cobry, „Thermoelectric Energy Harvesting from Transient Ambient Temperature Gradients”, Journal of Electric Material, vol. 41, pp. 1653–1661, 2012, https://doi.org/10.1007/s11664-011-1894-4.

[11] S. Sojan, “A Comprehensive Review of Energy Harvesting Techniques and its Potential Applications”, International Journal of Computer Applications (0975 – 8887), vol. 139, No. 3, April 2016.

[12] D. Koester, “Thermal energy harvesting for distributed sensors”, 53, December 2011.

[13] D. M. Rowe, CRC Handbook of Thermoelectrics (Boca Raton, CRC Press), 1995.

[14] “Generic Energy Harvesting Adapter Module for TEG”, App. Note TIDU808, Texas instruments, March 2015.

[15] Linear Technology Corp., datasheet LTC3108. [Online]. Available from: https://www.analog.com/media/en/technical-documentation/data-sheets/LTC3108.pdf , [accessed May 2020].

Design and Characterization of a 60 GHz Low-Noise Amplifier in GaAs m-HEMT Technology for Radar Detection Systems

Pape Sanoussy Diao, Thierry Alves, Benoit Poussot and Martine Villegas

ESYCOM, Univ Gustave Eiffel, CNRS UMR 9007, F-77454 Marne-la-Vallée, France
Email: pape-sanoussy.diao@esiee.fr, thierry.alves@esiee.fr

***Abstract*—This paper addresses the design of a 60 GHz low-noise amplifier for radar detection systems. Based on an impulse architecture, the required characteristics of the amplifier are determined to improve the system performance, especially in terms of range. The choice of the design technology is based on a detailed comparative study. Then, the amplifier is designed in a 70 nm gallium-arsenide metamorphic high electron mobility transistor technology. It includes three-stages transistors with inductive degeneration for a suitable trade-off between gain and noise. Critical design points related to coupling phenomena are identified in the layout realization. To limit these coupling effects, a progressive optimization method is used. The optimized amplifier achieves a gain of 14.3 dB and a noise factor of 2.1 dB at 60 GHz. The simulated non-linear characteristics show an input 1 dB compression point $IP_{1dB} = -9.6$ dBm and an input third-order intercept point $IIP3 \approx -4$ dBm. A good impedance matching at the input ($S_{11} < -15.4$ dB) and the output ($S_{22} < -16.3$ dB) is obtained in the frequency band of interest. The designed circuit consumes a total direct current power of 13.5 mW and occupies an area of 1.47×1.0 mm^2. In addition, the sensitivity characterization of the amplifier to voltage biasing, temperature and input impedance variations shows a good robustness of the design.**

Keywords–LNA design and characterization; 60 GHz; GaAs m-HEMT; Radar detection systems; Millimeter-wave technology.

I. Introduction

A 60 GHz Low-Noise Amplifier (LNA) has been concisely proposed in [1] for multi-band impulse detection system. But, for a more complete study, background, technology choice and design methodology need to be more detailed. In addition, the design characterization should be done taking into account the architecture considered and the targeted performance. This present study is motivated by the aforementioned reasons to bring a further development of the work presented in [1].

Availability of unlicensed bandwidth around 60 GHz in several regions of the world (57-66 GHz in Europe [2]) generates great interest in the development of new wireless systems. In radar, the use of millimeter-wave frequencies is often preferred to facilitate the detection of small objects (for example, bullets whose largest dimension does not exceed 10 cm). Furthermore, large bandwidths allow to achieve high spatial resolutions (a few cm). But, the use of millimeter-wave frequencies faces challenges related to the design, realization and even packaging of components. However, in recent decades, standards developments in millimeter-wave bands have contributed to significant advances in terms of integrated circuit technologies. For example, new process in Silicon-Germanium (SiGe) [3] [4] and III-V materials [5] [6] technologies allow the production of devices and systems for a variety of applications in millimeter-wave bands. These opportunities create a need for increasingly high-performance devices. In this context, this study adresses the design of a LNA with moderate gain and low power consumption, which has relatively wide bandwidth and small size in order to improve the performance of radar detection systems.

More generally, this work is part of a large study on the development of an ultra-wideband millimeter-wave detection system combining simplicity in terms of architecture and signal processing, low power consumption and small size. The objective is then to detect metal objects such as cylinders and plates, over short ranges (a few meters). In this study, the considered detection context is the monostatic radar configuration as described in Figure 1, where the target is a cylindrical metal of radius r and height h $(r; h)$. The detection system is schematized by a transceiver (TX-RX) using the same antenna, but it can also use two co-located antennas. The incidence angle θ indicates the orientation of the target with respect to the antenna boresight.

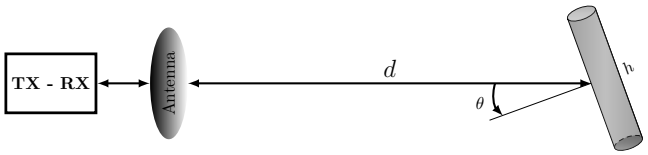


Figure 1. Context of the detection.

This contribution proposes the design and characterization of a LNA in a 70 nm Gallium-Arsenide (GaAs) metamorphic High Electron Mobility Transistor (m-HEMT) technology. It is organized as follows: Section II presents the detection principle and the associated architecture. In Section III the sizing of the proposed system is presented and the specifications of the LNAs are determined. Based on a detailed comparison, the design technology is chosen and described in Section IV. Then, the design principle is presented in Section V. In Section VI, we present the layout realization and electromagnetic simulations results. Design optimization and characterization are highlighted in Section VII. In addition, results are discussed and a comparison with the state-of-the-art is presented. And finally, a conclusion and future work are addressed in Section VIII.

II. Principle of the Detection

The angular dependence of the Radar Cross Section (RCS) of typical targets such as cylinders and plates (see Figure 2)

often results in a limitation of the detection range, particularly when they rotate away from the normal incidence. RCSs shown in Figure 2 are obtained from an electromagnetic simulation software called High Frequency Structure Simulator (HFSS). This limitation of the detection range occurs even when targets are close to the radar system and depends on their dimensions according to the operating wavelength [7]. To overcome this limitation, we use frequency diversity [8]. The idea is to take advantage of the frequency channel diversity due to the frequency dependence of the target RCS. This dependency is illustrated in Figure 3 with RCS simulations from HFSS for the normal incidence. Note that this frequency dependence of the RCSs may be more or less strong depending on the angle of incidence.

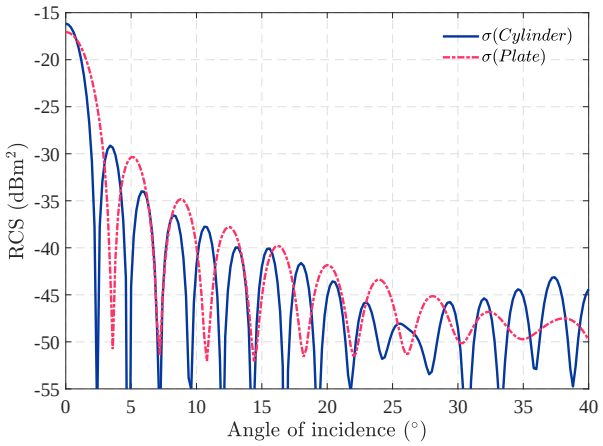


Figure 2. RCS of metallic targets as a function of elevation angle for cylinder (0.6 cm; 6 cm) and azimut angle for plate (0.5×4 cm^2) at 60 GHz.

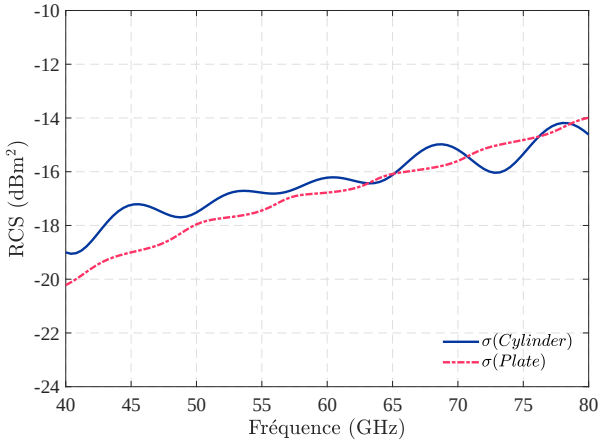


Figure 3. RCS of metallic targets as a function of frequency for cylinder (0.6 cm; 6 cm) and plate (0.5×4 cm^2) in normal incidence.

The proposed detection principle is then based on the impulse technique, using a multi-band approach [7] [9], to improve detection coverage, particularly the continuity of detection according to the target orientation angle θ. Frequencies around 60 GHz are chosen for spectrum availability [2], and also for their short wavelengths (5 mm at 60 GHz) to detect small objects, i.e., those whose largest dimension $\leq$ 10 cm.

The architecture associated with the detection principle is shown in Figure 4 in the case of a dual-band system around frequencies 57.8 GHz and 62.8 GHz. In the transmitter, Differential Structure Power Amplifiers (DSPAs), whose operation is based on the even mode rejection are used. DSPA$_1$ provides both signal division and pre-amplification. The signals selected by the filter bank are then amplified by DSPA$_2$ and transmitted by the same antenna. In reception, simple LNAs are used for better noise performance. The received signals are selected by a filter bank identical to the one in transmission before being subjected to the detection and decision process. More details of the architecture operation are given in [9].

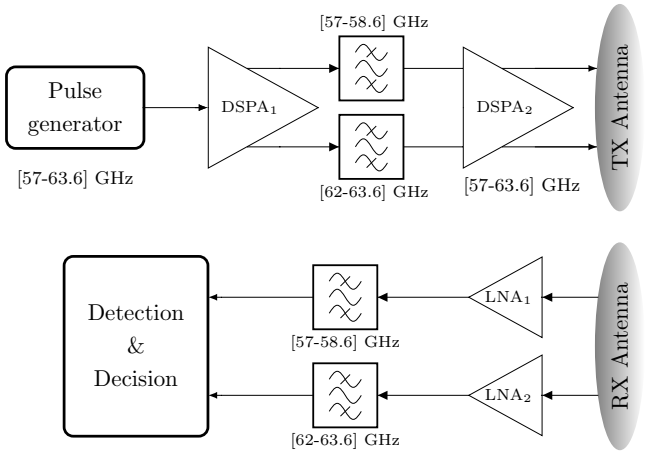


Figure 4. Dual-band detection architecture.

The proposed architecture can be applied in a more general case with N frequency bands. It has advantages in terms of simplicity of implementation by using impulse technique and does not required any frequency conversion. The processing of the received signals can be done by different techniques: selection combining, cumulative detection [9], coherent integration or non-coherent integration [10]. All of these techniques can help to improve the performance in terms of detection range compared to a single frequency band system.

III. Required Specifications of the LNA

To determine the specifications of the LNA, a full system sizing must be performed. For this, we start from the monostatic radar equation giving the maximum detection range (R_{max}) as a function of the minimum detectable signal power S_{min} [11]:

$$R_{max}^4 = \frac{P_t G_A^2 \lambda^2 \sigma}{(4\pi)^3 S_{min}} \quad (1)$$

where P_t is the transmitted power, G_A is the antenna gain, λ the operating wavelength and σ the RCS of the target.

As S_{min} is related to the thermal noise power of the receiver and the required Signal-to-Noise Ratio (SNR_r) to detect a target, the radar equation (1) can be written as follows:

$$R_{max}^4 = \frac{P_t G_A^2 \lambda^2 \sigma}{(4\pi)^3 \cdot kT\Delta f \cdot F \cdot SNR_r} \quad (2)$$

where k is the Boltzmann constant, T the noise temperature in K, Δf the receiver bandwidth in Hz and F its noise factor.

It is important to note that SNR_r is the minimum signal-to-noise ratio at the output of the receiver to ensure the detection of a given target.

In equation (2), the transmitted power and the antenna gain are determined by standardization [2]. The wavelength λ is chosen according to the application and the dimensions of the targets to be detected. The receiver bandwidth Δf is defined by the bandwidth of the front-end filters, which will set the range resolution ΔR of the system ($\Delta R = c/2\Delta f$). The SNR_r is defined by the desired performance in terms of detection and false alarm probabilities. The proposed detection principle is based on frequency and angle variations of the RCS. Thus, the receiver noise factor F is the only adjustable parameter to maximize system performance. Due to the position of the LNA in the receiver chain (see Figure 4), the impact of its noise factor will be more significant over that of the total Radio-Frequency (RF) chain according to the Friis formula:

$$F = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + ... + \frac{F_n - 1}{G_1 G_2 ... G_{n-1}} \quad (3)$$

where F_i and G_i are the noise factor and the power gain, respectively of the i-th stage, and n is the number of stages.

By setting the objective to detect a cylindrical metal target ($r = 0.6$ cm; $h = 5.4$ cm), up to 2 m at normal incidence, we will determine the required characteristics of the receiver stage and in particular those of the LNA. For this purpose, we extend our system to four bands in 57-66 GHz, distributed around the frequencies 57.8 GHz, 60.2 GHz, 62.8 GHz and 65.2 GHz. The output power of each channel of the $DSPA_2$ is set at 15 dBm (taking into account frequency bands standardization) and the gain of the antennas G_A at 12 dBi (which can be achieved with 4 patches of 6 dBi each). The bandwidth of each frequency channel is 1.6 GHz and the SNR_r depends on the detection technique used. Filter losses are set at 3.5 dB by referring to [12]. In a conventional single-band architecture, the SNR_r to ensure the detection of a nonfluctuating target with a detection probability of 90% and a false alarm probability of 10^{-6} is 13.2 dB [11]. In the case of a multi-band architecture, with a non-coherent integration of 4 pulses from different frequency bands sufficiently spaced, the SNR_r is only 8.3 dB for the same detection and false alarm probabilities. Based on this case of non-coherent integration and using equation (2), we established the technical specifications of the LNA given in Table I, to ensure the targeted detection objective. Reflection coefficients are chosen < -10 dB for an efficient power transfer at input and output of the LNA, otherwise a higher gain will be required.

TABLE I. TECHNICAL SPECIFICATIONS OF THE LNA

Parameters	*Values*
Bandwidth BW	$\geq$ 1.6 GHz
Power gain G	$\geq$ 12.5 dB
Noise figure NF	$<$ 3 dB
S_{11} & S_{22}	< -10 dB

In addition, to determine the predominant parameter of the LNA on the system performance, we studied the influence of its Gain (G) and Noise Figure (NF) on the detection range. This study presented in Table II evaluates the range variations R_V as a function of the variations in G and NF of the LNA. It is based on the following equation obtained by inserting (3) in equation (2):

$$R_{max}^4 = \frac{P_t G_A^2 \lambda^2 \sigma}{(4\pi)^3 \cdot kT\Delta f \cdot SNR_r \cdot \left(NF + \frac{F_f - 1}{G}\right)} \quad (4)$$

where F_f represents the filter losses according to the detection architecture shown in Figure 4.

TABLE II. LNA G AND NF INFLUENCE ON SYSTEM RANGE

G (dB)	*NF (dB)*	*R_V (%)*	*NF (dB)*	*G (dB)*	*R_V (%)*
11-14	2	1.1	2-5	11	16.8
11-14	3	0.9	2-5	12	17.9
11-14	4	0.7	2-5	13	18.2
11-14	5	0.6	2-5	14	18.3

The results show that, on average, a variation of 50% in the noise figure leads to a variation of about 18% in the detection range, while a same variation of 50% in the gain influences only about 1% the range of the system. This demonstrates the overriding of noise figure over gain in improving detection performance particularly in terms of range, as it can be directly seen in equation (4). This result will be very helpful in the choice of the design technology to realize the LNA.

IV. CHOICE OF THE DESIGN TECHNOLOGY

We have two design technologies for the realization of the LNA: SG13S, which is a Bipolar-Complementary Metal Oxide Semiconductor (BiCMOS) SiGe from IHP [13] and D007IH, an m-HEMT GaAs from OMMIC [14]. In order to choose the most suitable technology for the system requirements, we compare them in two steps. The first step focuses on passive elements such as transmission lines (TLs), inductors and capacitors. In the second step, a study of a simple amplifier designed with both technologies is proposed.

The realization of inductances and capacitances becomes more complex as the frequency rises, partly due to increased losses in dielectric substrates and conductors, but also due to coupling effects, which become significant. In order to take these problems into account, we have compared the losses of the TLs and the quality factor (Q-factor) of the grounded TLs, inductors and capacitors. Figure 5 shows a comparison of the losses of the two technologies based on simulations with Advanced Design System (ADS) from Keysight and HFSS. With SG13S technology, Metal1, which is at the lowest metallization level, is chosen as the ground plane. TLs and inductors are made with TopMetal2, which is located at the top level. With D007IH technology, the ground plane is located below the chip. TLs and inductors are made with the metal IN, which is on the top level. The results show a good fit between the ADS and HFSS simulations. They also clearly show that losses are higher with the SG13S (about 0.4 dB/mm more than the D007IH at 60 GHz). Indeed, in the SG13S technology model, there are no losses in the substrate (tangent loss $\tan\delta = 0$). This implies that losses come mainly from conductors. In fact, the ground plane on SG13S is too thin (0.42 μm) compared to 3.5 μm for D007IH. This partly explains the fact that the metal thickness is higher in the SG13S with 3 μm compared to 1.25 μm for D007IH to avoid too much losses. Furthermore, the substrate thickness between the ground plane and the metallization level of the TLs in

the D007IH is nearly ten times greater than the thickness of the substrate between Metal1 and TopMetal2 with SG13S. This results in narrow tracks to make 50 Ω TLs with the SG13S. These different characteristics explain why the losses are higher with SG13S. Finally, this is confirmed by the negative reactance of the TLs characteristic impedance higher in absolute value in SG13S than in D007IH (not shown here).

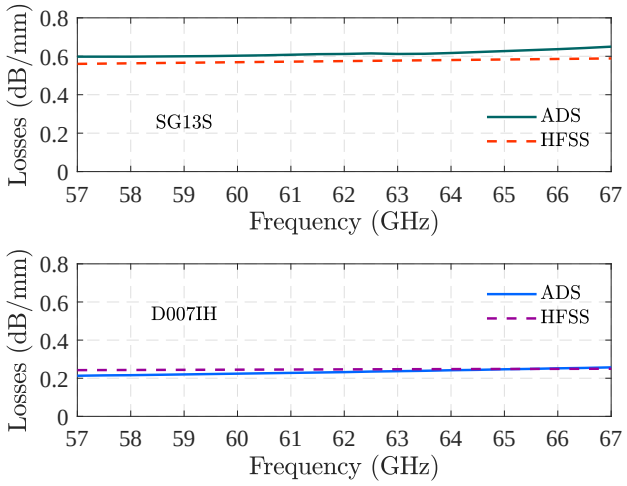


Figure 5. Comparison of 50 Ω TL losses.

A comparison of grounded TLs with the same characteristic impedance of 75 Ω giving rise to an inductance of 0.27 nH at 60 GHz is shown in Figure 6. D007IH exhibits a Q-factor more than 5 times higher than SG13S at 60 GHz.

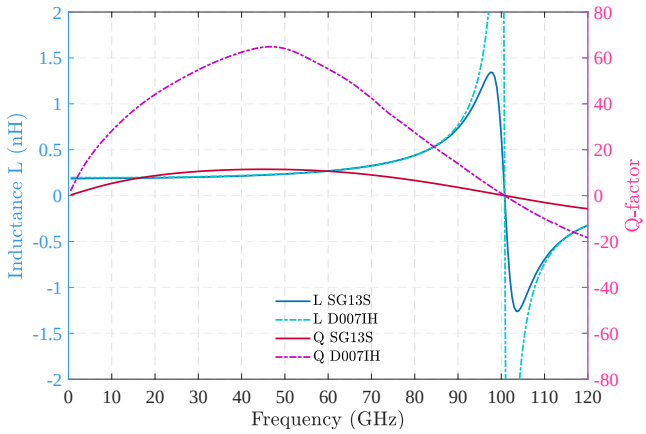


Figure 6. Comparison of grounded TLs.

TABLE III. PASSIVE ELEMENTS PERFORMANCES AT 60 GHz

Parameters	*SG13S*	*D007IH*
50 Ω TL Losses (dB/mm)	0.6	0.22
75 Ω grounded TL Q factor	11	56
1 nH inductance Q factor	14	11
250 fF capacitance Q factor	13	28

The performances of passive elements of the two technologies, based on electromagnetic simulations with ADS Keysight, are summarized in Table III. For the same coil inductor value of 1 nH at 60 GHz, the two technologies have almost equal Q-factor. However, it should be noted that inductors in SG13S are less integrable than those in D007IH because of their shape. For the same capacitor of about 250 fF at 60 GHz, the D007IH has a Q-factor of twice times higher that of the SG13S. On the other hand, for an equivalent surface, the density of a capacitor is higher with SG13S than with D007IH. However, the maximum capacitance value with SG13S is limited to only 8 pF, whereas with D007IH it can achieve 50 pF. As resistors are concerned, we note that the resistance achievable with SG13S are much higher than those possible with D007IH. Resistivities are generally higher on silicon (Si) than on GaAs.

Generally, GaAs technologies have better loss performance than Si technologies. This is partly due to the fact that the resistivity of GaAs substrates is higher than that of Si. Given the characteristics of the passive elements, the SG13S has the advantage of facilitating the realization of more integrable components (lines, capacitors, resistors). On the other hand, it presents too high losses for low noise applications. Unlike SG13S, the D007IH technology has the advantage of having low losses. At 60 GHz, the use of TLs is often preferred for the realization of inductors in particular. Therefore, the D007IH technology, due to its low losses, presents a major advantage for low noise applications. However, at millimeter-wave frequencies, the performance of a circuit does not depend only on the performance of the passive or active elements taken separately. Therefore, an overall analysis of the full circuit elements is necessary. For this reason, we have evaluated the performance of single-stage amplifiers designed with both technologies based essentially on distributed elements and for an operation at 61.5 GHz (see Table IV). The amplifier with

TABLE IV. PERFORMANCE OF SINGLE STAGE AMPLIFIERS

Technology	*Transistor*	*G (dB)*	*NF (dB)*	*P_{DC} (mW)*
SG13S	Bipolar	4.0	2.6	2.0
SG13S	MOS	3.1	2.6	9.2
D007IH	m-HEMT	3.9	1.2	4.1

MOS transistor is the least efficient in terms of gain, NF and power consumption (P_{DC}). The one with bipolar transistor has advantages in terms of gain and power consumption, but its noise figure is high. With the m-HEMT transistor, the amplifier has a gain almost equal to that of the bipolar transistor, with moderate power consumption (more than half that of the MOS transistor), and a noise figure of 1.4 dB lower than that of the bipolar and MOS transistors. Although its reverse isolation (S_{12}) is lower than those of bipolar and MOS transistors, the use of a multi-stage structure will allow to achieve a good level of isolation. Thus, performance with the m-HEMT transistor, with its low noise level, relatively good gain and moderate power consumption seems more suitable to meet our system requirements.

Based on the comparisons made in Tables III and IV, it is clear that the D007IH technology is better suited than the SG13S to meet the required technical specifications set out in Table I, particularly in terms of noise. This is supported by the fact that, regarding to the design, the influence of the LNA noise figure is more significant than that of its gain on the system range. Nevertheless, in order to satisfy the required gain level, the use of a multi-stage structure will be unavoidable.

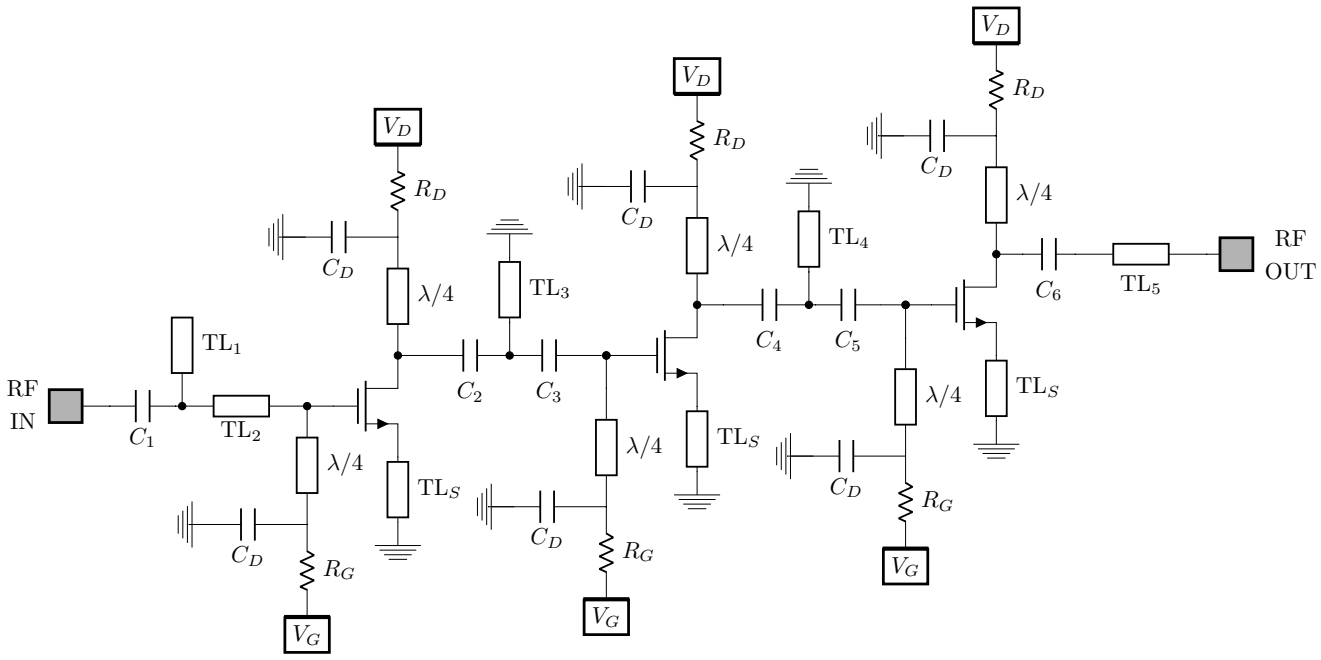


Figure 7. Schematic of the three stages LNA.

D007IH is a 70 nm gate length GaAs technology providing f_T/f_{max} of 300 GHz/450 GHz. It offers a depletion transistor m-HEMT [15], with high transconductance of $g_{m_{max}} = 1600$ mS/mm, which can support a voltage $V_{DS_{max}}$ of 3 V and a maximum current $I_{DSS_{max}}$ of 400 mA/μm. This type of transistor offers good performance in terms of noise, with a noise factor of only 0.5 dB at 30 GHz, giving it a privilege for security applications (millimeter-band imaging), telecommunications or radars. The process of the technology consists of a 3.5 μm metal in its underside, a 100 μm thick GaAs substrate above which different metallization levels are distinguished. These metallization levels are separated by dielectric layers of silicon-nitride (SiN) and silicon-dioxide (SiO_2). The most used metal layer for TLs realization is named IN metal and has a thickness of 1.25 μm. It is also possible to associate this layer with a gold layer of the same thickness for less losses.

V. Design Description of the LNA

The design of the LNA is done with ADS Keysight. Taking into account various simulation tests, the common source topology has been retained. Indeed, this topology is more suitable for low noise components compared to the cascode and common grid topologies [16], the latter offering better stability, higher gain, but also a not desired higher noise figure. In addition, the common source topology is simple to implement and provides moderate gain. In order to meet the required technical specifications of the LNAs established in Table I, particularly in terms of gain, we used a three-stages structure, as shown in Figure 7 [17] [18]. To optimize the performance of an amplifier, transistors size and bias point must be properly chosen. The size of the transistors is selected so as to ensure a good trade-off between gain and noise for the optimal bias point. A parametric study allowed to choose a transistor with 2×25 μm gate development and a voltage $V_{DS} = 1$ V.

The stability of the LNA is ensured by making each stage unconditionally stable. As the transistors are identical for all stages, they are degenerated in the same way so that the overall circuit is unconditionally stable in a larger bandwidth. Figure 8 shows the stability of the first stage through the Rollet's stability factor (K) and the parameter B. In addition, a resistor is used in the bias circuits to further stabilize the LNA. The value of this resistor is meticulously chosen because it influences the gain and noise performance of the LNA [19]. With the voltage drop across the resistor, the V_D potential is increased to 1.1 V to ensure a current $I_{DS} = 4.1$ mA. Moreover, the degeneration of the source allows to closer the Maximum Available Gain (MAG) and NF circles together, as shown in Figure 9. This makes the input stage matching easier [20].

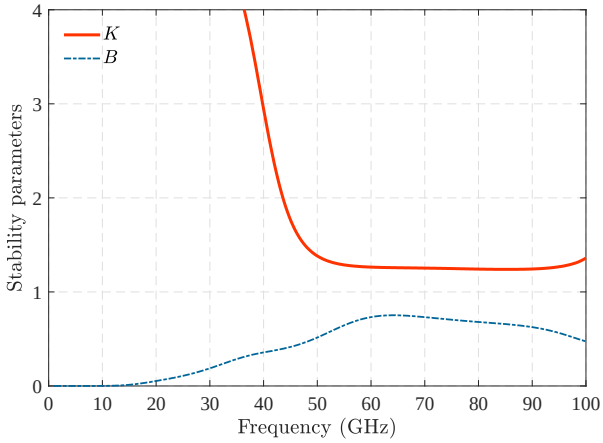


Figure 8. First stage stability.

In Figure 7, each bias circuit is made with a quarter-wave TL, bypass capacitor (C_D) to redirect RF leakage to ground,

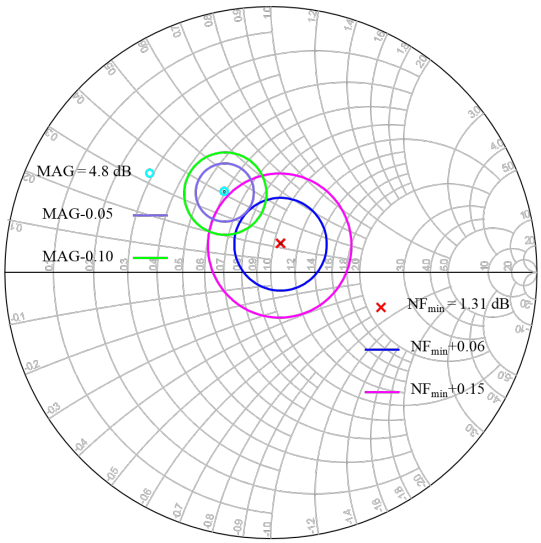


Figure 9. Gain and noise circles for input matching.

and a GaAs implanted resistor (R_G or R_D) to improve low frequency stability. The input matching network is formed by the link capacitor C_1, the open transmission line TL_1, which acts as a capacitor and an inductor made by TL_2. The inter-stages matching networks 1-2 and 2-3 are formed by identical topologies C_2, TL_3, C_3 and C_4, TL_4, C_5 respectively. This topology also makes it possible to isolate the power supplies of the different stages. The output matching is made by a simple serial structure C_6 and TL_5, optimized for a better gain, without too much degradation of the total noise figure of the LNA.

VI. Layout Realization and Simulations Results

At high frequencies and particularly at 60 GHz, the differences between schematic and layout simulations are often significant depending on the accuracy of the passive components models, which is related to the technology. Thus, to minimize these differences, elements such as bias circuits, Direct Current (DC) and RF pads, transistor access and link capacitors have been fixed from their electromagnetic models. This approach allowed the design to be lighter, since only the matching networks will then be optimized during the transition to the full layout.

The layout realization is done progressively in two main steps. Firstly, by considering the results of the partial electromagnetic simulations of the components i.e., each element is simulated alone without taking into account the others. And secondly, the overall simulation of the full LNA circuit. In addition, it should be noted that progressive optimizations have been made in order to maximise the overall performance of the circuit. The resulting 60.2 GHz LNA layout is shown in Figure 10 including RF and DC pads. All TLs are made with the same metal layer denoted IN and ground connections are realized using octagonal vias holes. The size of the circuit is 1.29×1.56 mm^2. Figures 11 and 12 present a comparison between the partial (p-index) and global (g-index) simulations of the LNA. Note that all layout simulations are done with ADS Momentum Microwave solver.

The results show differences related to coupling phenomena between the different elements of the circuit, which are not

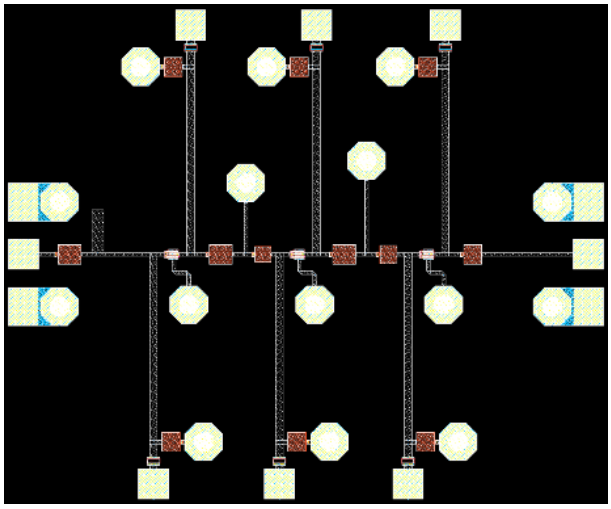


Figure 10. First layout of the LNA.

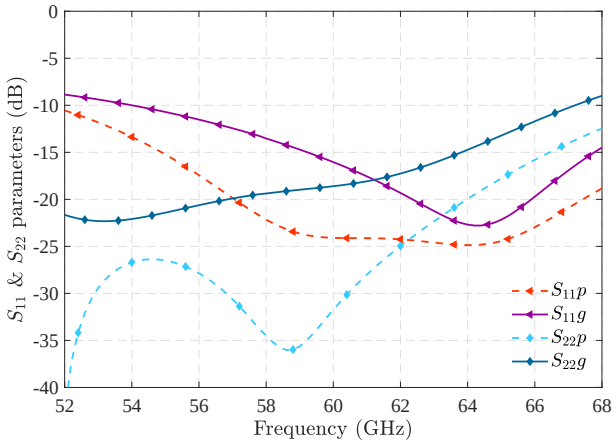


Figure 11. S_{11} and S_{22} with partial and global layout simulations.

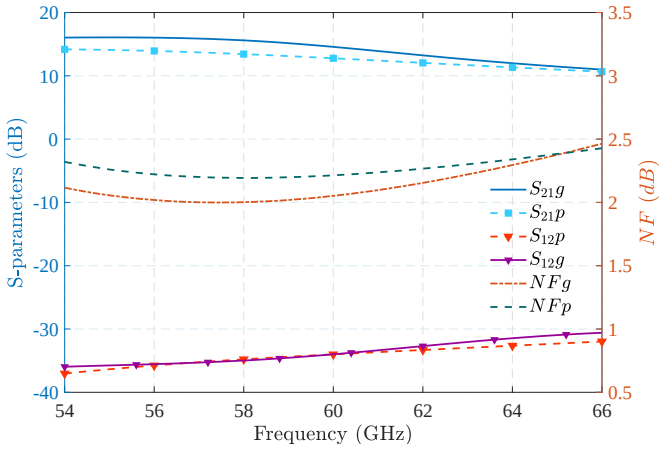


Figure 12. S_{21}, S_{12} and NF with partial and global layout simulations.

taken into account in the partial simulations. These couplings result in a more or less pronounced degradation of the circuit performance. In our case, we can see on Figure 11 the rise in reflection coefficients at both input S_{11} and output S_{22}, but also a slight improvement in the S_{21} coefficient and the noise figure,

as can be seen in Figure 12. In practice, it is very difficult to estimate these couplings. However, a re-optimization has been done to minimize the reflection coefficient degradations, but also to improve as well as possible the gain and noise performances. Moreover, the stability of the circuit is perfectly ensured at least until 100 GHz as shown in Figure 13.

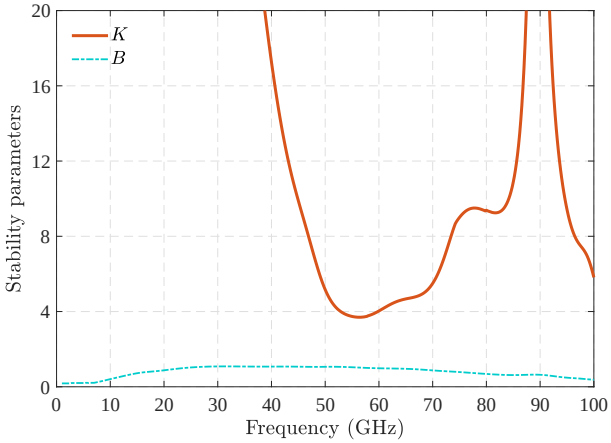


Figure 13. Stability parameters K and B of the whole LNA.

The global simulations results meet the targeted specifications. A gain of 14.5 dB and a noise figure of 2.06 dB are noted at 60.2 GHz. Reverse isolation is approximately -34 dB and the reflexion coefficients S_{11} and S_{22} are less than -15 dB in the band of interest (59.4-61 GHz). The -3 dB bandwidth (for $S_{11} < -10$ dB and $S_{22} < -10$ dB) ranges from 54 to 65 GHz with a fluctuation of about 0.38 dB in noise figure. In the band of interest, the variation in gain is about 0.9 dB and the noise figure ripple is less than 0.1 dB. In addition, the LNA consumes about 13.5 mW.

VII. Design Optimization and Characterization

To gain in miniaturization, we have optimized the size of the LNA. This is done progressively in two main steps. We started by reducing the length of the bias circuits so that the LNA does not exceed 1 mm following to the height of Figure 10. This was done by replacing the straight quarter-wave TLs by equivalent meandered lines. The use of meandered lines more or less changes the impedances presented at the input and output of the transistors. Therefore, the matching networks have been re-optimized. Thus, the capacitors and the grounded TLs have been resized in order to reduce the overall size of the circuit while minimizing the potential coupling between circuit elements. This made it possible to limit the rise in coefficients (S_{11} and S_{22}) and to guarantee the required performance in terms of gain and noise. The optimized LNA is shown in Figure 14 and occupies an area of 1.0×1.47 mm^2. Its S-parameters and noise figure are shown in Figure 15 and the non-linear characteristics in Figures 16 and 17.

Table V presents a comparison of the LNA performance before and after optimization in the band of interest for the detection system. With a gain of 14.3 dB and a noise figure of 2.1 dB at 60.2 GHz, the optimized LNA has almost identical performances with the one shown in Figure 10. The reflexion coefficients are less than -15.4 dB for S_{11} and -16.3 dB for S_{22} in the band of interest. Also note that the stability of the LNA

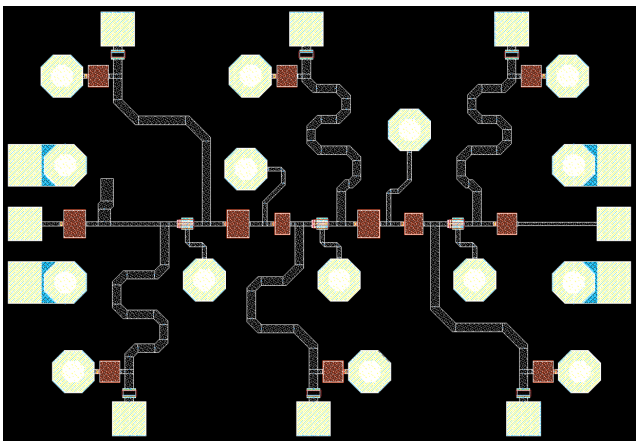


Figure 14. Layout of the optimized LNA.

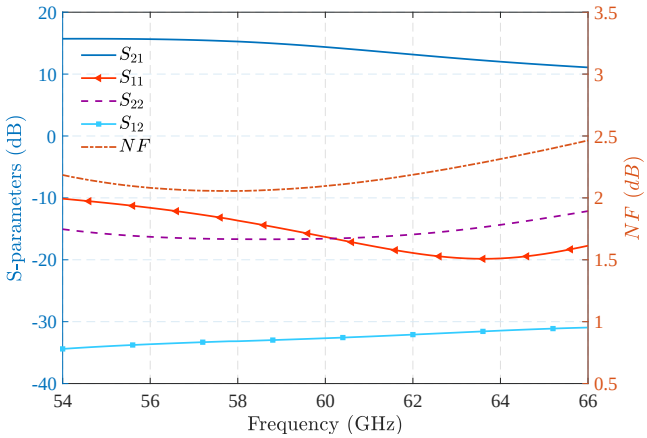


Figure 15. S-parameters and noise figure.

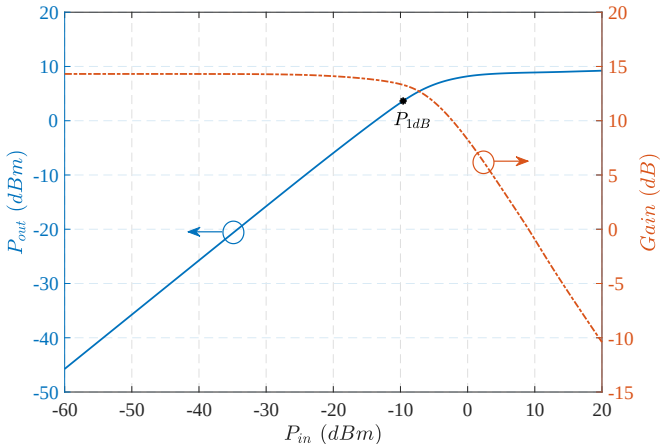


Figure 16. Gain and 1 dB compression point.

has been verified, and as before optimization, unconditional stability is ensured up to at least 100 GHz. The non-linear characteristics show good performances with a very slight improvement form -10 dBm to -9.6 dBm for the IP_{1dB} and from -4.4 dBm to -3.9 dBm for the $IIP3$. All of the above mentioned results prove a meticulous optimization of the LNA with an area reduction of about 27% , while maintaning the same power consumption of 13.5 mW.

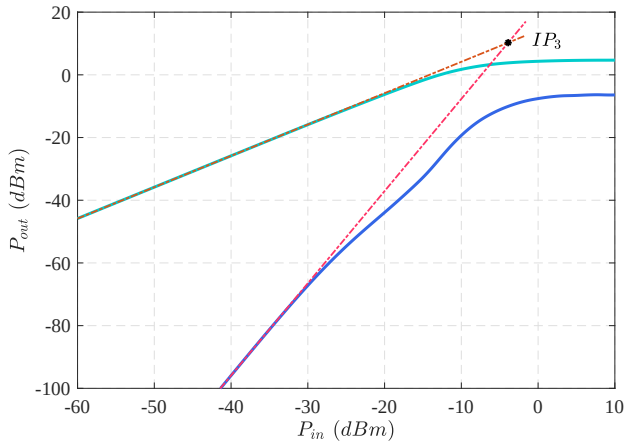


Figure 17. Third order interception point (IP3).

TABLE V. FEATURES OF THE OPTIMIZED LNA IN 59.4-61 GHz

Parameters	*First LNA*	*Optimized LNA*
S_{21} (dB)	14.5 ± 0.045	14.3 ± 0.4
NF (dB)	2.06 ± 0.03	2.10 ± 0.03
S_{11} (dB)	< -15.2	< -15.4
S_{22} (dB)	< -18	< -16.3
S_{12} (dB)	< -33.3	< -32.4
IP_{1dB} (dBm)	-10	-9.6
$IIP3$ (dBm)	-4.4	-3.9
Area (mm^2)	1.29×1.56	1.0×1.47

To further characterize our design, we are now evaluating its sensitivity to bias voltages, temperature and load variations in the band of interest (59.4-61 GHz). These studies are necessary to see how these parameters will influence the performance of the designed LNA and more globally the performance of the system.

Voltage sensitivity is evaluated by quantifying the variations in S_{21} and NF following the bias voltages V_{DS} and V_{GS} variations respectively, depending on whether their influences are more or less significant. Results are shown in Figures 18 and 19 with a variation of ± 10 % around the nominal bias voltage. The results of these studies show that our circuit is not very sensitive to bias voltage errors. Indeed, a 20% variation in the V_{GS} voltage leads to a variation of about 2 dB in gain

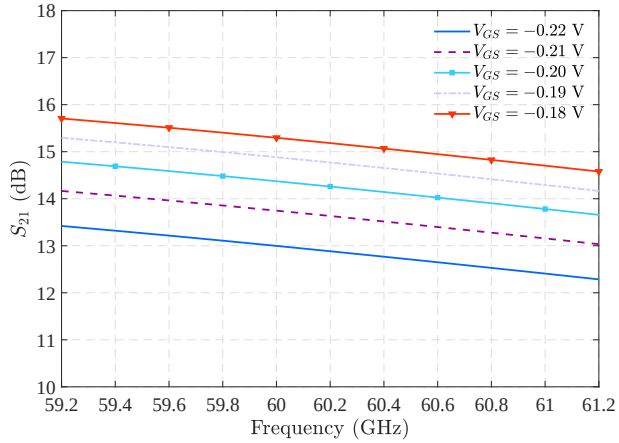


Figure 18. S_{21} sensitivity to bias voltage.

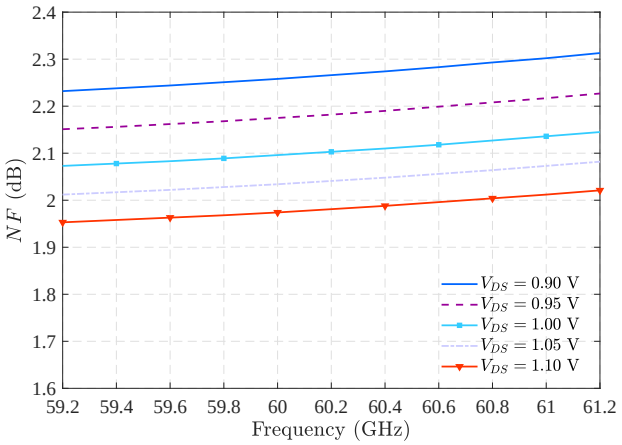


Figure 19. NF sensitivity to bias voltage.

(< 50 %), which has a negligible influence on the range of the system (see Table II). On the other hand, a 20% variation of the V_{DS} leads to a fluctuation of less than 0.3 dB in the noise figure. This corresponds to a variation of less than 1%, thus a negligible influence on the system range as it can be deduced from Table II.

Usually, the simulation temperature is taken at $16.85°$ C (290 K) as recommended by the IEEE standard for noise figure measurements. Thus, to characterize the temperature sensitivity of the LNA, we have simulated it under different temperatures ranging from $-40°$ C to $125°$ C. The results obtained in noise figure variations are shown in Figure 20. Gain variations are not presented here, because they are obviously negligible (< 0.1 dB). The results show a maximum variation of 0.24 dB in the noise figure between 59.4-61 GHz. As with the voltage sensitivity, these results show a negligible influence of noise figure variations on the system range when the LNA is subjected to large temperature variations.

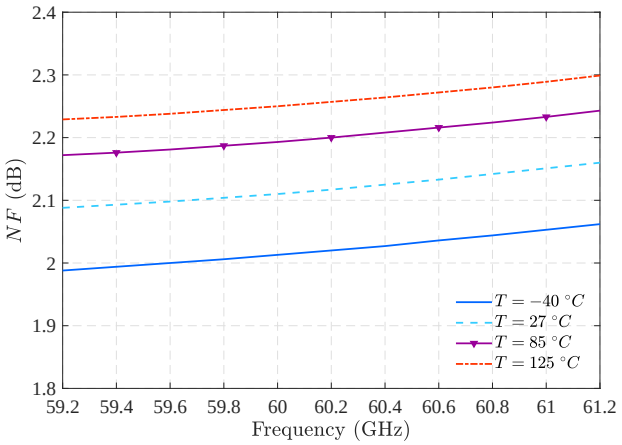


Figure 20. NF sensitivity to temperature.

After voltage and temperature, it seems particularly significant in the configuration of the detection architecture to study the influence of the impedance mismatch at the LNA input. In other words, it means seeing the influence of the antenna impedance with respect to that of the LNA input. To do this, we evaluated the variations in gain and noise figure of the LNA

TABLE VI. PERFORMANCE COMPARISON

Ref.	*Technology*	*Freq. (GHz)*	*G (dB)*	*NF (dB)*	*IP_{1dB} (dBm)*	*IIP3 (dBm)*	*P_{DC} (mW)*	*FoM_N*	Area (mm^2)
[4]	130 nm SiGe BiCMOS*	57-66	20.5	4.3	-17.8	-11,1	9.8	0.51	0.41 x 0.32
[19]	100 nm GaAs m-HEMT$^+$	60-90	19	2.5	-	-	56	0.20	3.5 x 1.0
[21]	40 nm CMOS$^+$	60	12.5	3.8	-	-15	20.4	0.30	0.63 x 0.31
[22]	65 nm CMOS$^+$	60	23	4	-26	-	8	0.70	0.35 x 0.14
[23]	90 nm CMOS$^+$	58-77	11.2	4.8	-18.7	-7.4	10	0.41	0.72 x 0.76
[24]	65 nm CMOS$^+$	60	20.2	5.2	-25	-	28	0.13	0.54 x 0.80
[25]	65 nm CMOS*	61	22	5.5	-	-10.7	26	0.13	0.71 x 0.46
[26]	130 nm SiGe BiCMOS+	52-56	15	3.3-3.6	-13.5	-	19.6	0.37	0.30 x 0.35
[27]	50 nm GaAs m-HEMT$^+$	60-90	27	2.6	-26	-	45	0.23	1.6 x 2.3
This work	**70 nm GaAs m-HEMT***	**60**	**14.3**	**2.1**	**-9.6**	**-3.9**	**13.5**	**1.0**	**1.47 x 1.0**

$^+$ Measures
$*$ Simulations

by changing its input impedance with a $\pm 10\%$ variation around the nominal value of 50 Ω, as shown in Figure 21. The results show that both the gain and the noise figure have a fluctuation of less than 0.1 dB when the input impedance of the LNA is varied from 45 Ω to 55 Ω. Again, the performance of the LNA shows negligible variations that will not have significant influence on the detection system range.

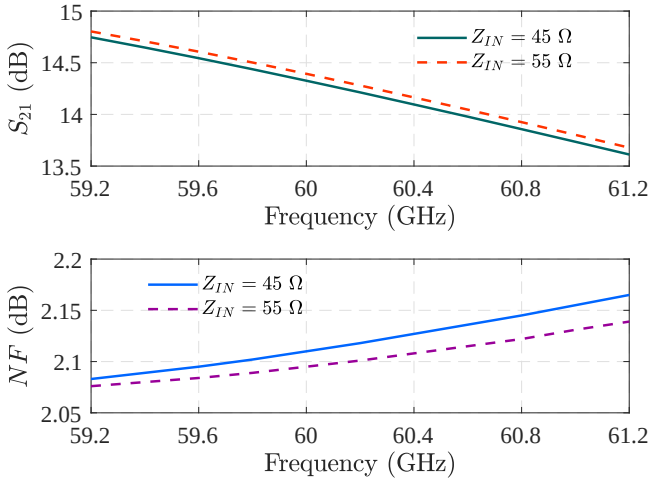


Figure 21. S_{21} and NF sensitivity to input impedance Z_{IN}.

Concisely, sensitivity studies reveal a high robustness of the designed LNA with respect to variations in bias voltage, temperature and input load (mismatch). In fact, in all three cases, the maximum variation in gain is about 2 dB for a variation of $\pm 10\%$ in V_{GS} voltage, while for the noise figure it is less than 0.3 dB. These sensitivity performances avoid a significant decrease of the system range when the LNA is subjected to significant variations in voltage, temperature or input load.

Finally, the performance of the designed LNA compared with the state-of-the-art is summarized in Table VI. LNAs designed around 60 GHz with different technologies are considered. Usually, the figure of merit (FoM) of an LNA is calculated as a function of the gain, noise figure and power consumption as defined in [3]. But, considering that the influence of the gain is negligible on the detection performance of the system (particularly in terms of range, as shown in Table II), we define the FoM as follows:

$$FoM = \frac{1}{(NF - 1) \cdot P_{DC}} \tag{5}$$

This definition of the FoM is more consistent to our detection context, especially in relation to the proposed architecture. Furthermore, it should be noted that in Table VI, the FoM_N is the normalized FoM to our LNA result. Regarding Table VI, our LNA shows good performance, especially in terms of noise figure, even if some of the results are based on measurements. With a moderate power consumption compared to other III-V technologies [19] and [27] or even the 40 nm CMOS [21], it presents a good gain and meet the targeted detection objectives. Note that more gain can be achieved by increasing the drain voltage of the output stage or adding a fourth transistor. But this would obviously increase the power consumption and a little more the noise figure. In addition, it is clear that the non-linear characteristics of our designed LNA are much better than those of the m-HEMT and CMOS technologies presented in Table VI. As an example, its IP_{1dB} is -9.6 dBm, while it is less than -17 dBm for [4] [22]- [24] and [27]. The same is observed for the $IIP3$. Moreover, our LNA is more integrable compared to the reported same type technologies (50 and 100 nm GaAs m-HEMT). In our detection context, our design is more efficient with a better figure of merit.

In summary, the performance of the designed LNA satisfy perfectly the specifications established in Table I. Furthermore, with these performances and taking into account the proposed system (see Figure 4), the detection range is improved up to 2.3 m for the considered metal cylinder ($r = 0.6$ cm; $h = 5.4$ cm) at normal incidence ($\theta = 0°$), when a non-coherent integration is performed over the received signals in the four frequency bands. This represents an improvement of 30% in range compared to a conventional single-band detection system.

VIII. Conclusion and Future Work

This work focused on the design of a 60 GHz LNA for small metal objects detection system. Firstly, the detection principle was presented with an impulse architecture based on frequency channel diversity. Starting from the radar equation and targeted detection objectives, the critical components of the architecture were identified and the required LNAs specifications established. Following this, the design technology was choosen based on a comparative study between BiCMOS SiGe and m-HEMT GaAs. Then, the design was done in a 70 nm GaAs m-HEMT technology from OMMIC. Three stages transistors with inductive degeneration and bias circuits including resistance were used to better scale the transistor with a good trade-off between gain and noise, while ensuring unconditionnal stability up to 100 GHz. The LNA is completed with ADS Keysight. Critical design issues related to coupling phenomena were highlighted. Post-layout simulation results with ADS Momentum Microwave solver show good performance, especially in terms of noise. With a noise figure of 2.1 dB at 60 GHz, our LNA is much better that those commonly found in the state-of-the-art. For a moderate power consumption of 13.5 mW, which is relativeley low for III-V's technologies, it exhibits 14.3 dB of gain at 60 GHz. The reflection coefficients are less than -10 dB in the 54-68 GHz band. The LNA shows an input power at 1 dB compression point of -9.6 dBm and an input third order interception point of -3.9 dBm. Furthermore, the designed LNA occupies less space with 1.0×1.47 mm^2 compared to others of the same type of technology. In addition, it has been characterized in terms of sensitivity to voltage biasing, temperature and input impedance variations. This characterization shows good robustness of the design. At the end, it was compared to other recently published 60 GHz LNAs, particularly with a specific figure of merit adapted to our detection context. The results of our design show the potential of III-V's technologies, especially the 70 nm GaAs m-HEMT for very low noise applications, particularly to improve the performance of detection systems.

In terms of design, some future work is planned. First, we have the manufacturing and characterization by measurements. And in a longer term, the joint design of the reception chain (i.e., LNA, filter and detector) to further improve the system performances.

References

[1] P. S. Diao, T. Alves, B. Poussot, and M. Villegas, "60 GHz Low-Noise Amplifier in a 70 nm GaAs m-HEMT Technology for Multi-band Impulse Detection System," in Proceedings of the 12th International Conference on Advances in Circuits, Electronics and Micro-electronics (CENICS) Oct. 27–31, 2019, Nice, France. IARIA, 2019, pp. 19–24.

[2] M. G. L. Frecassetti, "E-Band and V-Band-Survey on status of world-wide regulation," ETSI White Paper, no. 9, June 2015, pp. 1–40.

[3] A. C. Ulusoy et al., "A SiGe D-Band Low-Noise Amplifier Utilizing Gain-Boosting Technique," IEEE Microwave and Wireless Components Letters, vol. 25, no. 1, 2015, pp. 61–63.

[4] M. Pallesen, "Design of a 60 GHz Low Noise Amplifier in a 0.13 μm SiGe BiCMOS Process," Master's thesis, The University of Bergen, 2016, URL: http://bora.uib.no/handle/1956/12595 [accessed: 2017-10-24].

[5] A. Dyskin, D. Ritter, and I. Kallfass, "Ultra wideband cascaded low noise amplifier implemented in 100-nm GaAs metamorphic-HEMT technology," in Proceedings of the International Symposium on Signals, Systems, and Electronics (ISSSE) Oct. 3–5, 2012, Potsdam, Germany. IEEE, Dec. 2012, pp. 1–4.

[6] Y. Chen et al., "OMMIC 70 nm mHEMT LNA design," in Proceedings of the IEEE Asia Pacific Microwave Conference (APMC) Nov. 13–16, 2014, Kuala Lumpar, Malaysia. IEEE, Jan. 2018, pp. 1192–1195.

[7] P. S. Diao, T. Alves, B. Poussot, and M. Villegas, "A new method and transceiver architecture dedicated to continuous detection of very small metallic object," in Proceedings of the 10th Global Symposium on Millimeter-Waves (GSMM) May 24–26, 2017, Hong Kong, China. IEEE, Jul. 2017, pp. 169–171.

[8] D. K. Barton, Frequency Diversity Theory. Artech House Inc., 1977, vol. 6 of Radars, section 2, pp. 35–114, in Frequency Agility and Diversity, ISBN: 0-89006-067-3.

[9] P. S. Diao, T. Alves, M. Villegas, and B. Poussot, "Compact millimeter wave architecture dedicated to object detection using dual band-dual polarization and impulse method," in Proceedings of the 13th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME) June 12–15, 2017, Giardini Naxos, Italy. IEEE, Jul. 2017, pp. 161–164.

[10] P. Surendran, J.-H. Lee, and S. J. Ko, Performance of Non-coherent Detectors for Ultra Wide Band Short Range Radar in Automobile Applications. Springer-Verlag Berlin Heidelberg, 2012, vol. 377, pp. 185–195 in Software Engineering Research, Management and Applications 2011, ISBN: 978-3-642-23201-5.

[11] M. I. Skolnik, The Radar Equation, 2nd ed. McGraw Hill, Inc., 1980, chapter 2, pp. 15–67, in Introduction to Radar Systems, ISBN: 0-07-057909-1.

[12] R. Abdaoui, M. Villegas, G. Baudoin, and A. Diet, "Microstrip band pass filter bank for 60 GHz UWB impulse radio multi band architectures," in Proceedings of the IEEE MTT-S International Microwave Workshop Series on Millimeter Wave Integration Technologies Sep. 15–16, 2011, Sitges, Spain. IEEE, Oct. 2011, pp. 192–195.

[13] "SG13S Process Specification Rev. 1.06," July 2016, URL: https://www.ihp-microelectronics.com/en/services/mpw-prototyping/sigec-bicmos-technologies.html [accessed: 2017-10-18].

[14] "D007IH Design Manual - OM-CI/008/MG," Oct. 2017, URL: http://www.ommic.fr/site/mpw-4r [accessed: 2018-07-20].

[15] M. Ney, HEMT's capability for millimeter wave applications. Hermes Penton Science, 2002, chapter 2, pp. 20–42, in Millimeter Waves in Communication Systems, ISBN: 19039 9617 1.

[16] T. Das, "Practical Considerations for Low Noise Amplifier Design," RFLNA White Paper Rev. 0, Freescale Semiconductor, Inc., May 2013, pp. 1–9.

[17] P. S. Diao, T. Alves, B. Poussot, and M. Villegas, "A 60 GHz Low-Noise Amplifier for Detection Systems," in Proceedings of the 2019 IEEE Radio and Antenna Days of the Indian Ocean (RADIO) Sep. 23–26, 2019, Reunion. IEEE, Jan. 2020, pp. 1–2.

[18] P. S. Diao, "60 GHz Low-Noise Amplifier for Detection Systems," IOP Conference Series: Materials Science and Engineering, vol. 766, Mar. 2020, p. 012003.

[19] A. Bessemoulin, J. Grunenputt, P. Felton, A. Tessmann, and E. Kohn, "Coplanar W-band low noise amplifier MMIC using 100-nm gate-length GaAs PHEMTs," in Proceedings of the 34th European Microwave Conference Oct. 12–14, 2004, Amsterdam, The Netherlands, vol. 1. IEEE, 2005, pp. 25–28.

[20] S. P. Voinigescu et al., "A scalable high-frequency noise model for bipolar transistors with application to optimal transistor sizing for low-noise amplifier design," IEEE Journal of Solid-State Circuits, vol. 32, no. 9, 1997, pp. 1430–1439.

[21] H. Gao et al., "A 48–61 GHz LNA in 40-nm CMOS with 3.6 dB minimum NF employing a metal slotting method," in Proceedings of the IEEE Radio Frequency Integrated Circuits Symposium (RFIC) May 22–24, 2016, San Francisco, CA, USA. IEEE, Jul. 2016, pp. 154–157.

[22] E. Cohen, O. Degani, and D. Ritter, "A wideband gain-boosting 8 mW LNA with 23 dB gain and 4 dB NF in 65 nm CMOS process for 60 GHz applications," in Proceedings of the IEEE Radio Frequency Integrated Circuits Symposium June 17–19, 2012, Montreal, QC, Canada. IEEE, Jul. 2012, pp. 207–210.

[23] Y.-S. Lin, C.-Y. Lee, and C.-C. Chen, "A 9.99 mW low-noise amplifier for 60 GHz WPAN system and 77 GHz automobile radar system in 90 nm CMOS," in Proceedings of the IEEE Radio and Wireless Symposium (RWS) Jan. 25–28, 2015, San Diego, CA, USA. IEEE, 2015, pp. 65–67.

[24] C. So and S. Hong, "60 GHz variable gain LNA with small NF variation," in Proceedings of the IEEE International Symposium on Radio-Frequency Integration Technology (RFIT) 30 Aug.–1 Sep., 2017, Seoul, South Korea. IEEE, Sep. 2017, pp. 171–173.

[25] A. Wang, L. Li, and T. Cui, "A transformer neutralization based 60 GHz LNA in 65 nm LP CMOS with 22 dB gain and 5.5 dB NF," in Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS2013) May 19–23, 2013, Beijing, China. IEEE, Aug. 2012, pp. 1111–1114.

[26] S. Zihir and G. M. Rebeiz, "A wideband 60 GHz LNA with 3.3 dB minimum noise figure," in Proceedings of the IEEE MTT-S International Microwave Symposium (IMS) June 4–9, 2011, Honololu, HI, USA. IEEE, Oct. 2017, pp. 1969–1971.

[27] P. M. Smith et al., "A 50 nm mHEMT millimeter-wave MMIC LNA with wideband noise and gain performance," in Proceedings of the IEEE MTT-S International Microwave Symposium (IMS2014) June 1–6, 2014, Tampa, FL, USA. IEEE, Jul. 2014, pp. 1–4.

Real-time Evaluation of Failure and Reliability in Agricultural Sprayers Using Embedded Sensors and Controller Area Bus Protocol

Paulo E. Cruvinel[1,2], Heitor V. Mercaldi[1,3], Pedro B. Andrade[1], Elmer A. G. Penãloza [1,4]
[1]Embrapa Instrumentation Brazilian Agricultural Research Corporation, São Carlos, SP, Brazil
[2] Post-Graduation Programs in Computer Science - Federal University of São Carlos, SP, Brazil
[3]Department of Electrical Engineering – Federal University of São Carlos, SP, Brazil
[4]Engineering Center - Federal University of Pelotas, RS, Brazil
Emails: paulo.cruvinel@embrapa.br, heitor@ufscar.br, pedro_borghi@hotmail.com, eagpenaloza@ufpel.edu.br

Abstract—**This paper presents a method for the operational analysis of agricultural sprayers based on their smart sensor devices. It is becoming increasingly necessary to study a sprayer's reliability, which is one of the major concerns in real-world agricultural machinery during field operation. A method is thus presented in this study for the verification of the reliability and indication of the failure of an agricultural spraying system using smart sensors, a microcontroller, and a controller area network protocol for communication and data analysis. Smart sensors were used for sensing the pressure, flow of pesticides, and temperature for the spray quality control evaluation. These smart sensors play an important role in supporting variable control, and they should not only be operating correctly but should also ensure the verification of application quality, which depends on the correct rate of pesticide application for pest control. Furthermore, these smart sensors were embedded in the main parts of the sprayer machine. Such a system facilitates the real-time and low-cost periodic verification of the sensors' calibration as well as the evaluation of the entire operation, in addition to indicating necessary corrections and sensor replacements. Such an innovative system would play a strategic role in allowing users to appropriate such knowledge and decrease the measurement errors in variables that are directly related to pest-control efficiency as well as reduce the resulting impact on the environment.**

Keywords—Real-time processing; Failure and reliability; Calibration of sensors; Agricultural sprayers; Decision-making support; CAN bus; Precision farming.

I. INTRODUCTION

Embedded sensor systems are computer-based systems that can be part of larger systems. The former perform some of the requirements of these larger systems. The majority of such embedded systems are also characterized as real-time systems. These systems, especially in agriculture, are required to meet stringent specifications for safety, reliability, availability, and other attributes of dependability [1]-[3].

The high complexity of embedded real-time systems results in increasing demands with respect to engineering requirements, high-level design, early error detection, productivity, integration, verification, and maintenance, which increases the importance of the efficient management of its life-cycle properties, such as maintainability, portability, and adaptability [4].

The constructive design of dependable distributed real-time systems using pre-validated components requires precise interface specifications of the components in the temporal and value domains [5]. The digital transformation and its impacts on agricultural automation have been providing, from the technological point of view, tools for risk-management development based on the use of the agriculture 4.0 concepts [6]–[9]. Such an approach facilitates the rational use of agricultural inputs and the promotion of paths for realizing improved productivity and sustainability gains.

Today, not only in scientific research areas but also in the agricultural industry, it has become possible to use scalable computational architectures, mainly those based on embedded and smart ones [10]. Such architectures have the potential to comprise multiple processor nodes with the use of language to allow the implementation of new integrated risk models. Without agricultural mechanization and its advanced automation, it will be practically impossible to meet such needs and provide solutions for realizing food and nutrition safety [11]-[15].

The automation of agricultural machinery is an intensive area of research and development with an emphasis on the enhancement of food quality, preservation of operator comfort and safety, precision application of agrochemicals, energy conservation, and environmental control. Current automation applications are oriented towards and assist in the attainment of environment friendly and more sustainable systems of agricultural and food production [16][17]. The global mechanization of farming practices has revolutionized food production, thus enabling it to keep pace with the global population growth [18][19].

In terms of the current technology development for agriculture, there is a need for more investments, system innovation, and a better understanding of how people and machines can interact to each other. In addition, almost every piece of agricultural equipment comprises sensors and controls these days, and a number of sensing technologies are used in agriculture for providing data that help farmers monitor and optimize crop cultivation. In such a context, the assessment of sensor failure and reliability is important for machinery design engineers and researchers.

The methodology for assessing sensor reliability has adaptive aspects and should be customized as a function of the sensor application. In agriculture, for instance, the design of embedded electronic-sensor-based systems in machinery for field operation has been shown to require the inclusion of failure and reliability evaluations. Thus, three general approaches for designing such sensor-based systems can be considered: system failure owing to uncalibrated sensors, system-failure rate prediction, and physics-of-failure reliability assessment [20]-[23].

In general, the sensors in an agricultural sprayer are organized in sensor networks. If using the concept of redundancy, the failure of a single device may not be critical to the pesticides application. However, when failures occur in sensors, the consequences are likely to be disastrous, particularly in the case of critical applications, such as the application of pesticides for pest control. The impact of an incorrect application of pesticides is well known, not only in terms of the related economic aspects and the plant's health, but also its effect on the environment. The cause of such a failure must be determined as soon as possible; otherwise, the negative consequences could become more widespread [24].

The droplets sprayed during a pest control process are of different sized, and a percentage of the liquid volume is sprayed as fine droplets, regardless of the nozzle model used. However, the droplets should be of uniform size in order to realize the necessary efficiency for pest control. Similarly, some parameters are used to analyze the spectrum of the droplet size of sprays, such as the Volume Median Diameter (VMD) and the Coefficient of Variation (CV). The analysis of the spectrum uniformity comprise the use of the values of terms presented in $DV_{0.1}$, $DV_{0.5}$, and $DV_{0.9}$, which represent the droplet diameters such as the percentage of cumulative spray volume, which means, equal to 10%, 50% and 90% respectively. The diameter of 50% of the cumulative volume ($DV_{0.5}$) represents the VMD of a pesticide spray's application.

The uniformity of the droplet distribution is not only dependent on internal machinery parameters but also on external factors. The internal parameters are related to the sprayer machinery, i.e., the temperature and pressure of the syrup in the sprayer's boom, as well its flow in the nozzles [25]. The external factors are mainly related to the strength of the wind and its direction as well as the ambient humidity and temperature [26]. Although the internal factors can be controlled, the external ones cannot be controlled, and the latter contribute to the drift effect, which is undesirable. Nevertheless, such an effect can be minimized by selecting the type of nozzle based on the type of application required for the pest control management of an agricultural crop [27].

However, it should be noted that the internal sprayer's variables play an important role for the efficiency of the pest control process and should be correctly adjusted and controlled. Two terms that serve to express uniformity are the Extension (E_x) and Relative Amplitude (RA), which quantifies the range that covers 80% of the spray volume, and a comparative index respectively. The higher the value of the RA, the more heterogeneous is the spray spectrum, which implies a lower distribution uniformity.

Figure 1 presents three different results of pesticide applications for pest control that have the same VMD but different uniformities and RA values [28][29]. It can be observed in Figure 1(a) that the application has good homogeneity as the sizes of the drops applied are very similar, thus indicating a low variation in the size of the drops. This indicates that this pesticide's application process is more effective in contrast to that presented in Figure 1(c), which shows very small drops and very large drops, which can affect the quality of the pesticide's application process. In Figure 1(b), an intermediate result is presented, which can also occur during the pesticide's application on a real agricultural field.

According to the records, isolated component evaluations of sprayers have been performed since the 1940s, but only in the 1970s did technical inspection programs emerge [30]. Around 1960, the implementation of the first Sprayer Inspection Project began in Germany. In 1969, other countries, such as Italy, began to perform inspections and, since then, the improvement of quality and the reduction of the negative impacts of these applications were observed [31]. There exist reports that state that agricultural sprayers have been inspected since 1991 in Norway [32]. The periodic inspection of sprayers implemented in Europe, in addition to the verification of reliability demonstrate the importance of the educational process [33]. In Belgium, obligatory inspections have been performed on agricultural sprayers in use since 1995, with the main objectives set as the maintenance of the equipment and the education of applicators [34][35]. In a project executed in Spain's Valencia region, the inspected sprayers were divided into operative or non-operative as a function of their condition of use [36]. In Argentina, a survey conducted in the 1990s indicated the need for the technical maintenance of spraying machines because the majority of them were malfunctioning [37].

In Brazil, the first sprayer inspection was performed in 1998, where an evaluation was performed in the State of Paraná, and inadequate working conditions of the pressure gauges of some of the sprayers were observed [38]. Today, in several countries, periodic sprayer inspections are performed, and various groups of researchers have reported that the best conditions for the use of sprayers are closely related to their constant maintenance. In such a context, the uniformity of the spray distribution realized by the sprayer boom, working pressure, temperature of the mixture, and volume of the pesticide, which should be adjusted for effective pest control [39]-[43], play important roles in realizing the best conditions for the use of sprayers.

Currently, agricultural spraying is used with a focus on precision agriculture, wherein control, supervision, and the highest quality of the application process are sought, to increase the safety and efficiency of the application processes. These aspects are also related to the minimization of the environmental impacts resulting from these agrochemical application processes. In work focused on the quantification of the economics of the localized application

(variable rate), it is common to observe improvements in the cost/benefit relation [44][45].

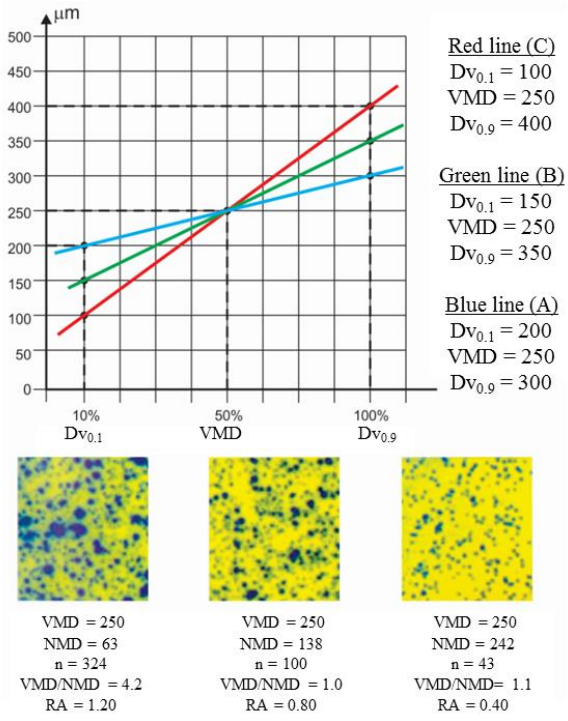


Figure 1. Examples of different uniformities observed for the pesticide's applications, with them all having the same VMD but with differences in the distribution and RA values. The values of $DV_{0.1}$, VMD, and $DV_{0.9}$ should be observed (adapted from [28] and [29]).

Variables, such as the temperature, flow, and pressure of the pesticide in a sprayer, have a direct influence on these results, thus affecting the volume and distribution of the drops in the plantation, which directly influence the efficiency of the application. If there is no control over volume and distribution of the pesticide drops, wastage of the pesticide can occur. Extremely fine drops can be carried by the wind, thus spreading and contaminating the environment, which characterizes the drift phenomenon. Extremely large droplets—although they reduce the occurrence of drift—provide less coverage of the application target because the pesticide volume that the leaves of crops can hold is limited by their size [46][47]. Therefore, it is important to know the precise values of the variables of temperature, pressure, and flow to have greater control over the application of these agricultural products. For the automation of those sprayers processes, embedded computer systems are currently being used.

The innovation presented in this report is based on the concept of on-the-go periodic measurements for operational surveillance based on the monitoring of the flow, pressure, and temperature of the mixture (pesticides plus water) to not only obtain, in real time, the information regarding the operational failure, but also the sprayer reliability analysis.

When performed properly, a periodic evaluation of the sensors' calibration or even a verification of the electronics used for signal processing can rectify mistakes, and network robustness can be established. In addition, the selection of a reliability assessment approach is of fundamental importance because it is related to the effective design of strategies for the operation of reliable sensors.

Furthermore, research on sensors and their effects on the reliability and response characteristics of agricultural sprayer devices during their operation are presented herein. The presented concept and the obtained results can be used in various sprayers' modalities and can make improving their reliability possible in relation to the sensor calibration, which defines the quality of the application of pesticides. As the control circuits rely on the feedback from voltage/current sensors, the performance of the whole system used for the pesticide application is likely to be affected by the sensors' failure rates, dynamic characteristics, and signal-processing circuits. This approach proactively incorporates reliability into the process by establishing a method of verifying the calibration of the sensors, i.e., including verification modules for important variables of the spraying process in an unsupervised and automated interface.

This work has been focused on the temporal specification of interfaces in composable distributed real-time systems, and four principles of composability have been established, which include the independent development of nodes, stability of prior services, performability of the communication system, and reduplication of the determinism.

This system presents the temporal firewall interface that forms a fully specified operational interface for failure and reliability evaluation in agricultural sprayers. This paper explains how the temporal firewall interface supports the four principles of composability. The interfaces are then classified from the point of view of composability, and how these interfaces correspond to the time-triggered and event-triggered communication paradigms is demonstrated.

After this introduction, there is Section II, which describes the materials and methods used in this study. In Section III, the results obtained are discussed, and the conclusions are presented in Section IV.

II. MATERIALS AND METHODS

To design the module to be developed for the virtual verification of the calibration of the sensors in a spraying system, the use of a low-cost Arduino architecture was considered. For the validation of the developed module, the platform developed at the Brazilian Agricultural Research Corporation (Embrapa Instrumentation) in partnership with the School of Engineering of São Carlos University of São Paulo (EESC-USP) was used [48]. This platform is used for sprayer development and performs analyses and operates as an agricultural sprayer development system (ASDS). It uses a National Instruments embedded controller, NI-cRIO, which works on the platform LabVIEW. The NI-cRIO architecture integrates four components: a real-time processor, a user-programmable field-programmable gate array, a modular input/output system and a complete software tool chain for programming applications. This ASDS is an advanced development system that makes

possible the design of architectures involving the connections of hydraulic components and devices, mechanical pumps, and electronic and computer algorithms. Such a system also comprises hydraulic devices used to develop any configuration of commercial agricultural sprays and new prototypes of sprayers, a user interface for system monitoring and control, and an electromechanical structure that emulates the movement of the agricultural sprayer in the field (Figure 2).

Figure 2. ASDS dedicated to the application of liquid agricultural inputs.

The ASDS platform comprises the following components: (1) spray nozzle, (2) system that emulates the movement of the sprayer, (3) pesticide disposal tank, (4) user interface for the development system, and (5) spray booms. In such a platform, the data are presented via a graphical user interface (GUI), where the user can interact with the digital devices via graphical elements with icons and visual indicators, thereby allowing them to select and manipulate symbols to obtain a practical result.

For the organization of a reference database comprising accurate reference values for flow, pressure, and temperature, calibrated and high-precision sensors were used. The sensors were subjected to known temperature, pressure, and flow conditions to obtain voltage values related to these conditions. In such a context, it was important to observe the droplet size (Table I), which influences the effectiveness of the spraying in covering the target and penetrating the leaves into a plant. Smaller droplets have a better coverage capacity, i.e., they offer a greater drops/cm^2 value). Furthermore, smaller droplets provide greater penetration capability and are recommended when good coverage and penetration are required. However, smaller droplets can be more sensitive to evaporation and drift processes. In productive agricultural systems, in general, large drops are preferred for the application of herbicides, such as glyphosate, while fine droplets are preferred in the case of insecticides, fungicides, and other products of less systemicity.

The extant technical literature comprises broad-nozzle descriptions, their recommended use, the selection of the proper nozzle type, and calibration method. However, any modification in the values of the temperature and pressure on the boom will change the drop characteristics, i.e., the sprayer´s operation.

Although the appropriate selection of the nozzle type is relevant, is also very important to take into consideration the technology for pesticides application. Therefore, the whole sprayer system is involved in the application process, not only in determining the amount of spray applied to an area, the uniformity, and the coverage, but also the target and the potential amount of drift. Furthermore, during operation, the nozzles facilitate the breaking of the mixture into droplets and also propel the droplets in the appropriate direction. Drift can be minimized by selecting the response time of the sprayers, the best time for applications with respect to climatic conditions, the controllers, which are used to obtain the optimal pressure or even the optimal volume, as well as the nozzle that produces the required droplet size while providing adequate coverage at the intended application rate.

TABLE I. SPRAY TIP CLASSIFICATION BY DROPLET SIZE (BASED ON THE STANDARD ASAE S-572)

Classification category	Symbol	Color code	Approximate VMD (μm)
Very Fine	VF	Red	<100
Fine	F	Orange	100-175
Medium	M	Yellow	175-250
Coarse	C	Blue	250-375
Very Coarse	VC	Green	375-450
Extremely Coarse	XC	White	>450

It is still important to observe and take into consideration that even when a tip predominantly produces large drops, there exists a small portion of fine droplets in the applied volume.

The controller area network (CAN) bus was also used. It is a synchronous serial communication protocol. Modules connected to a network send messages to the bus at known time intervals in order to realize the synchronization. The CAN bus was developed by Bosch [49] as a multimaster, message broadcast system that specifies a maximum signaling rate of 1 Mbps and wherein the modules can act as masters and slaves depending on their use [50]. This protocol works with multicast messages, wherein all the modules connected to a network receive all the sent messages. The connected modules check the status of the bus and determine whether another module of a higher priority is not sending messages; if this is observed, the module whose message has the lowest priority interrupts the transmission and allows the highest-priority message to be sent.

Communication in a CAN network, in version 2.0A, occurs through messages or frames, which can be of the following types: data frame, remote frame, error frame, or overload frame. Each type of frame has specific internal fields, which are relevant to the information to be sent. The data frame in a CAN 2.0A network (Figure 3), which is of interest in this work, consists of five main fields:

(1) The arbitration field includes the identifier (ID; 11 bits) used to identify the message and solve problems related to message collisions; in such a context the remote transmission request (RTR; 1 bit) indicates whether it is a data or remote frame.

(2) The control field includes the extended ID (IDE; 1 bit), which indicates whether the frame has an ID of 11 bits (standard) or and 29 bits (extended). Furthermore, it comprises the r0 (1 bit), which is reserved for future modifications, and the data length code (DLC; 4 bits), which is used to convey the number of bytes of the data frame.

(3) The data field is named Data (0–8 bytes) for the transmitted message.

(4) The cyclic redundancy check (CRC; 15 bits) field is used to detect transmission errors.

(5) The fields acknowledge slot (ACK; 1 bit) has the function of indicating the time for which the transmitter waits for the indication that some node in the network has received the frame successfully.

It should be noted that the message or frame also has signals for start of frame (SOF; 1 bit), which indicates the beginning of a transmission, and end of frame (EOF; 7 bits), which is used to convey the end of a transmission. In addition, two more delimiters are also available: one for the CRC and the other for the ACK, both having 1 bit each.

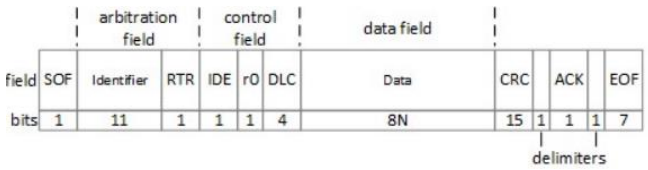


Figure 3. CAN 2.0A data frame for standard message identifier (11 bits). Both the ID and IDE bits are always 0.

The anti-collision mechanism based on the arbitration field, which allows the efficient exchange of messages in the CAN, implies additional attention for the network design. Such a requirement is to allow all "nodes" to transmit their messages at the desired time. Therefore, during the network design, it is necessary to consider the time for which each frame uses the CAN bus as well as the priority of each message or frame in order to avoid overloading the bus. In addition, with the increase in the network complexity and data traffic, it is necessary to analyze the Worst Case related to the Transmission Time (WCTT), which is the longest time gap between the queuing of a message and the arrival at the destination [51][52].

One of the main contributions of this work is a methodology for guaranteeing the validity of the variables measured using a calibration module. For such a purpose, a CAN network was implemented at a laboratory scale, wherein the traffic time in the network can be neglected in relation to the sample rates of the variables, thus ensuring the exchange of messages in real time. For this, the network bit rate was adjusted to 500 kbps and both the bus load and traffic time variation of each message in the bus were analyzed using the NI-CAN USB 8473 from National Instruments.

In networks of greater complexity, it is necessary to use other methodologies such as intelligent scheduling or dynamic ID allocation, and in the case of a control system, it may be necessary to compensate for the delay times [53].

The method used to obtain and approximate the model was based on the use of polynomial regression. In addition, such a concept can be used to estimate the expected value of a variable (y) given the value of another variable (x). This type of regression is used in models that obey polynomial and nonlinear behavior, as in the previous case. For these types of model, it is necessary to adjust for a higher-degree polynomial function [54]. This technique follows the same steps as those of linear regression but comprises the use of a concept based on Eq. (1).

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m + e \qquad (1)$$

where (y) is a polynomial function in which *a* represents real numbers (sometimes called the coefficients of the polynomial), *m* is the degree, and (*e*) represents an error. In this case, the mathematical procedure is the same as that used in the least-squares method, but the error is now represented by a function of degree greater than 1, as shown in Eq. (2).

$$\sum_{i=1}^{n} e_i^2 = \\ = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 + \cdots + a_m x_i^m\right)^2 \qquad (2)$$

Thus, Eq. (2) must be derived in parts from the terms that accompany x_i and be equated to zero in order to obtain a system of equations, which makes it possible to calculate its values.

Subsequent to the construction of a database comprising precalibrated values along with a mathematical model obtained, as mentioned earlier, an intelligent calibration and correction system can be applied using such a dataset as a reference to compare the results obtained from the sensors operating in real time. Such sensors have their calibrations checked periodically using the results obtained from the use of the models. Thus, using this comparison method, the system can determine whether a sensor is calibrated, i.e., the same concept can be replicated for each monitored variable. Furthermore, either a real-time recalibration can be performed or the sensor can be replaced, if necessary.

The methods of comparison comprise the use of the relative change, Euclidean distance (ED), Root Mean Square Error (RMSE), and percent error [55]. The ED and RMSE methods were used in the developed solution. For the comparison of the measured values of the variables, the ED takes into consideration the distance between two points that can be calculated via the application of the Pythagorean Theorem. In the algorithm, the ED is primarily calculated as

the square root of the sum of the squares of the arithmetic difference between the corresponding coordinates of two points, as shown in Eq. (3).

$$d(x,y) = \sqrt{(x - x_{ref})^2 + (y - y_{ref})^2} \quad (3)$$

where $d(x, y)$ is the ED, (x) is the measured point, (x_{ref}) is the measured variable at the reference point, (y) is the variable at the measured point, and (y_{ref}) is the variable obtained at the reference point. In addition, as a second verification, the RMSE is used. It represents the standard deviation of the residuals (prediction errors). The residuals are a measure of how far the data points are from the regression line, and based on such a concept the RMSE represents the standard deviation of the residuals (prediction errors). Thus, it indicates how concentrated the data are around the line of best fit and is given by Eq. (4).

$$RMSE = \sqrt{(x_{ref} - x)^2} \quad (4)$$

where (x_{ref}) is the reference variable, and (x) is the measurement variable.

In addition, an accurate power supply is used because such a system will be used not only for the verification of the sensors' calibration, but also their possible failure and reliability. For the calibration, it is necessary to consider one power supply that generates a precise and high-stability reference voltage. Such a voltage serves as a parameter for the intelligent calibration and correction system, and it is also responsible for feeding each of the electronic devices used.

The majority of analog-to-digital and digital-to-analog converters internally have voltage references that are used in the process of converting the signal, either to quantize its analog signal or to convert its digital signal to analog [56]. At this point, the accuracy and stability of the reference directly influences the conversion performance.

In agricultural spraying systems, the most commonly used sensors are as follows: (1) temperature sensors used to measure the temperature of the syrup, which is formed by the addition of the pesticides to water, as well as the temperature of the environment in which the spraying occurs; (2) pressure sensors used to measure the pressure in the spray bar near the spray nozzles; and (3) flow sensors, which measure the flow in the tubes and spray bar, and are used to measure and feed these values back to the spray quality control system.

The CAN controller comprises a hardware implementation of the CAN's protocol, and it is responsible for all the access control to the bus and for controlling not only the message arbitration but also the transmission rate. In this manner, the users are required to define the types of messages that will be sent and program the recorders.

The transceiver used was manufactured by NXP-Philips Semiconductors. It was used because it is fully ISO11898 compatible and supports high-speed CAN. It can also act as the entire interface between the network and the physical bus [57]. The inputs/outputs (pins 6 and 7) for the transceiver can be directly connected to the CAN L and CAN H lines of the CAN bus used. A 5-V power supply is applied at pin 2, and the ground potential (GND) is connected to pin 3. Pin 8 of the transceiver is called the "silent mode," and, if a 5-V voltage is applied at this pin, the mode is activated, thus preventing the component from sending CAN messages to the bus.

If no voltage is applied at this pin, the transceiver operates normally. Pin 5, named Voltage-Reference (VREF), provides the average CAN bus voltage, and pins 1 and 4, named TXD and RXD, respectively, are responsible for receiving or sending the serial signal that is used by the CAN controller to decode the CAN messages. At each decoded dominant bit, the transceiver sends a 1-bit serial via the TXD pin, and, at each recessive bit, the transceiver sends a 0 bit. In this manner, the messages are transferred bit-by-bit from the transceiver to the MCP2515 CAN controller, which decodes the sequence according to the CAN protocol.

Figure 4 presents both the CAN arrangement and how the communication between the transceiver and the microcontroller is performed.

The transceiver RXD pin receives the CAN message sent by the microcontroller. In addition, when a full message is received, it is passed to the CAN bus via the CAN H and CAN L pins.

The MCP2515, manufactured by Microchip, is a stand-alone CAN controller that implements the CAN specification, Version 2.0B. However, it is able to transmit messages in the CAN2.0A and B standards, that is, it can transmit and receive standard and extended data frames with 11 or 29 bits as message identifiers (frame IDs), respectively. The MCP 2515 was used because it makes the serial peripheral interface bus communication with another microcontroller possible, and its manufacturer, Microchip, provides the necessary instructions for writing and reading the registers. Each register has a byte for address that is used via some instructions to make the necessary settings. The addressing of each register is different from its content, that is, the initial setting of the bits for a register is not equal to the numerical value of its address

Figure 5 presents the integration of the Arduino-based architecture and a CAN with the sensors (temperature, pressure, and flow) in the sprayer system. The module that comprises the Arduino platform is a low-cost, functional, and easily programmable device. The Arduino Uno is a board consisting of an ATMEL ATMEG328 microcontroller and input and output circuits, and it can be easily connected to a computer via a universal serial bus cable and is programmed using free software called Arduino IDE (integrated development environment) and a language based on C/C++. The Analog-to-Digital (A/D) converter inputs, having a 10-bit resolution for a considered voltage range from 0–5 V, were used to read the signals related to the pressure and temperature sensors, which were obtained after the signals were correctly conditioned. The flow sensor used already has a digital output and its reading was obtained via timer 1 of the Arduino, which has a resolution of 16 bits.

Furthermore, it was configured as a counter, and in this manner, a reading was obtained at intervals of 50 ms.

Figure 6 depicts the software structure for the sensors' monitoring as well as the spraying process for the failure and reliability analysis. First, all sensors are tested in relation to failure. The process of monitoring the operation of the agricultural sprayer in real time is then started and is periodically repeated.

The flags are used to alert the operator of the operational status. Either the group of sensors or one of them can fail during an operation. For this reason and because of the probability of its occurrence, a previous routine is used to verify the operation based on the use of previously calibrated values and references of electrical voltage.

The reference modules receive an electrical signal from the Arduino architectures using the controller area bus protocol and determine whether they are calibrated or must be replaced.

As a function of the measured values of the variables, such as the flow, pressure, and temperature, a verification is performed periodically to determine whether the sprayer is operating adequately or if there is need for the adjustments of these variables. Such a verification can also indicate whether parts of the circuits related to each variable must be replaced when the correction of a failure cannot be made via the software. To obtain information regarding the operational conditions, a set of flags is used for signaling by the GUI. In addition, if a sensor is required to be recalibrated, the system performs the necessary correction to deliver the appropriate information to a CAN bus, where the control and processing unit collects the sensors' information of all the modules.

Furthermore, the CAN protocol has been used because its advantages involve data communication and the use of only two wires, which reduces its cost and facilitates its physical implementation.

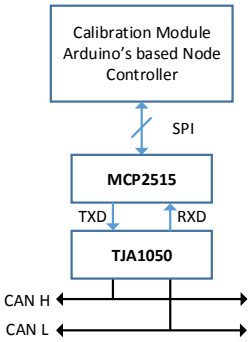
To communicate between the Arduino and the CAN bus, two important elements that are not directly found in the standard Arduino Uno were used. For this, a CAN transceiver (TJA1050) and CAN controller (MCP2515) dedicated to translating the signals made available serially by the transceiver were used [58].

III. RESULTS AND DISCUSSION

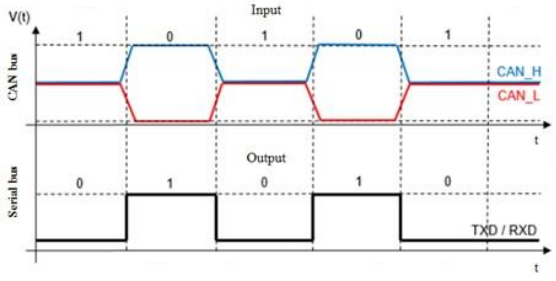
The sensors that generate analog signals were connected directly to the calibration module, which analyzed and corrected the data obtained through the algorithms. The data were then sent to the CAN network, which used the control and processing unit for presenting the information with the values calibrated by the supervision software. The implementation of the intelligent calibration and correction was performed using the Arduino-based architecture and the algorithms with the use of mathematical models. When the algorithm was started, it received the values of the sensor with the parameter to be analyzed, or temperature, pressure, or flow, and then this value was compared with those of the reference model constructed using the database. If the result of the comparison was satisfactory, this value was sent to the CAN bus; otherwise, this value was corrected by the software through emulation, and only then was the value sent to the bus. When the read values were outside the typical range of the sensors, there was an indication for sensor replacement, and the user was informed via a flag. There was a specific flag for each type of sensor, i.e., FLAG#1, FLAG#2, and FLAG#3, respectively, for the sensors used for the flow, pressure, and temperature measurements.

Reliability is an important performance index of agricultural sprayers. A paradigm shift in the reliability research of agricultural sprayers has resulted in the publication of a simple handbook based on a constant failure rate for the smart-system sensor-based and the support real-time decision-making approaches. Based on this, for each flag, the structure was considered to be that presented in Figure 7.

(a)



(b)



(c)

Figure 4. (a) Arduino CAN bus shield (MCP 2515); (b) block diagram of the structural architecture for the operation; and (c) input and output diagram of the TJA1050 transceiver.

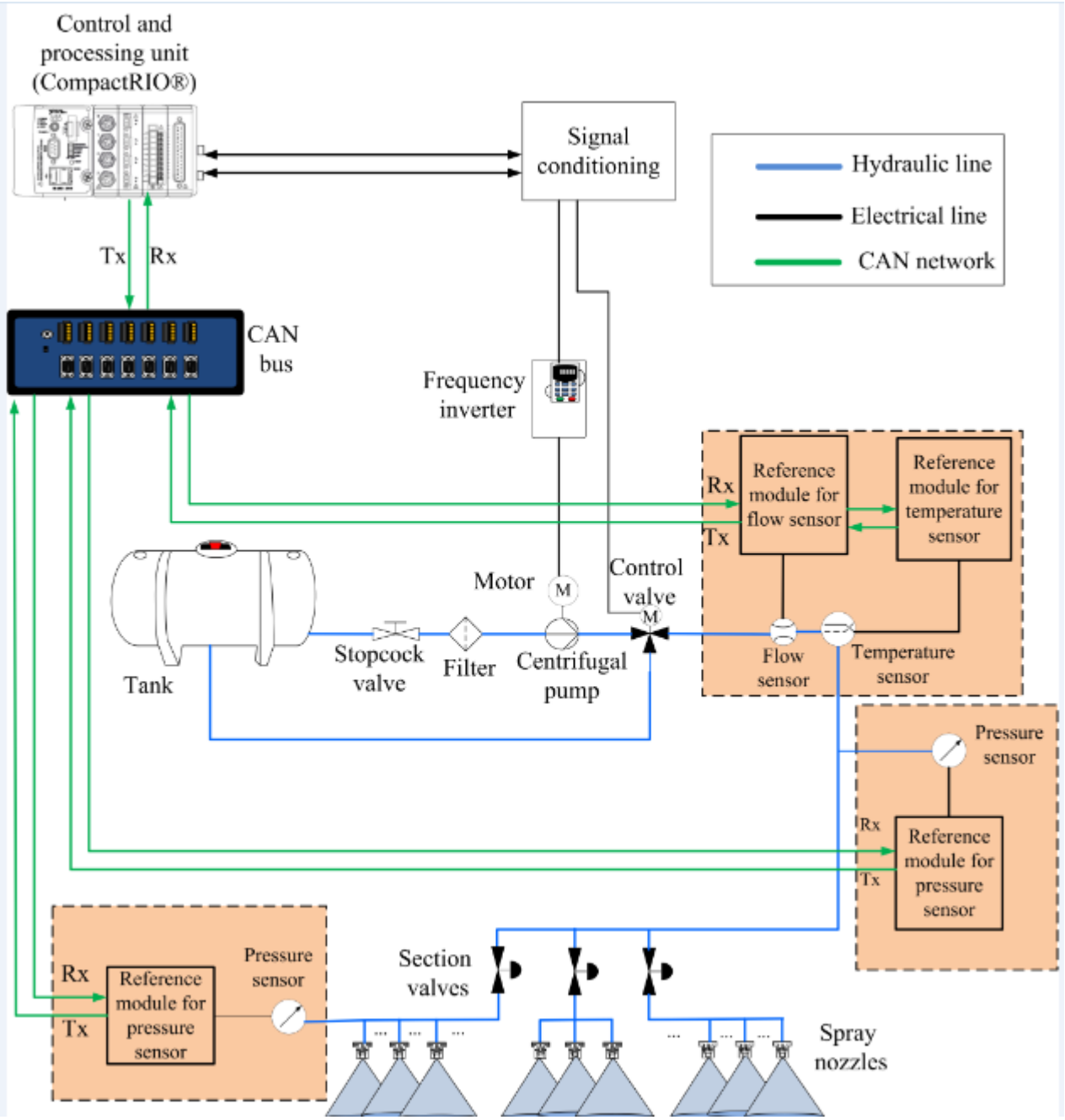


Figure 5. Block diagram of the sprayer system, in which the electrohydraulic configuration and the CAN network can be seen: in the red blocks are the modules based on the Arduino architecture, one for each sensor's modalities, for measurements of flow, pressure, and temperature.

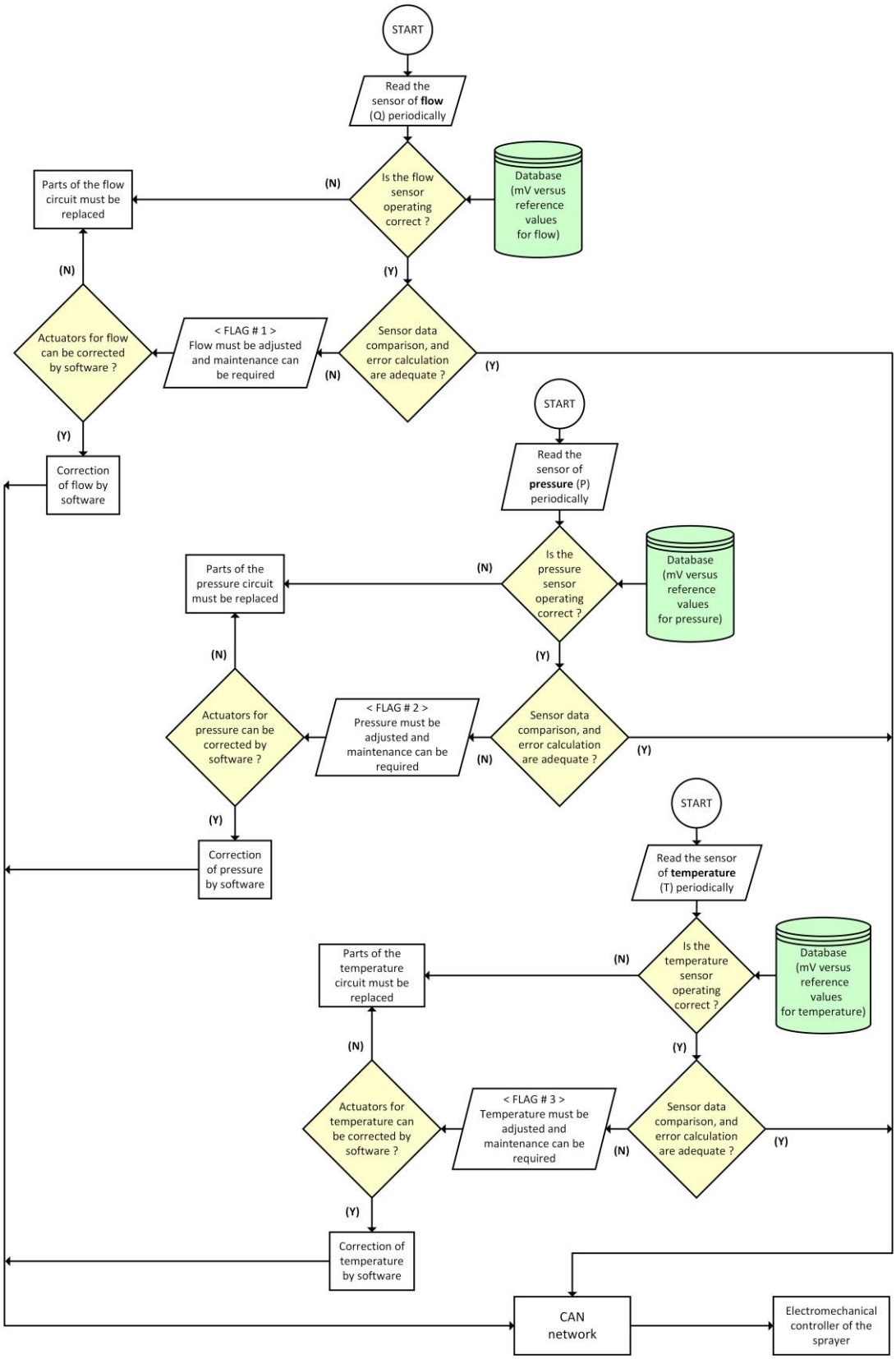


Figure 6. Software structure for monitoring the sensors used for measurements as well as the process used for spraying for the failure and reliability analyses.

For the flag structure, the context of sensitivity and specificity was used to summarize the performance of a diagnostic test with outcomes that determined the level of a standard for operation. When the test was quantitative, the receiver operating characteristic curves were used to display the performance of all the possible cut points of the quantitative diagnostic marker. Attention was focused on determining an optimal decision rule, which is also called the optimal operating point. Such a point provides a graphical interpretation for decision making. The construction of the databases for the three different sensors, which were related to the calibration and correct operation of the agricultural sprayer in a specific range of use, was organized in advance.

Table II lists the results as evidence of the successful operation of the algorithms applied for a commercially available sprayer's inspection based on the system for real-time failure and reliability analysis.

Safety and approval tests are used to find and guarantee that an approved safety element of an agricultural sprayer reliably or consistently functions in accordance with the manufacturer specifications. Furthermore, the robustness margin is based on the formulation of the robustness requirements in the agricultural industry. These tests do not require specific, detailed uncertainty models, and, hence, these margins can be evaluated based on the interpretation of the analyzed results. They are, in general, evaluated in the frequency domain, or even by using the information related to the safety margin of a machine's operation, without the loss of its hydraulic characteristics and purpose. Similarly, the design specifications and performance tests are typically related to the performance specifications. Besides, the specifications are written in projects and should be observed when implemented. The design specifications for a piece of machinery are straightforward in relation to its purpose and application.

Figure 7. Structure of the flags, in which the operational conditions of the sprayers based on the flow, pressure, and temperature as well as the constraints can be observed.

TABLE II. RESULTS FOR A REAL-TIME FAILURE AND RELIABILITY ANALYSIS.

FLAG # 1	**Destruction level (Q_6 and Q_7)**	**Design Specification/ Performance tests (Q_4 and Q_5)**	**Robustness margin (Q_2 and Q_3)**	**Safety and approval tests (Q_1)**	**Optimal operating point (Q_0)**
[l/m]	$3.00 \leq F_6 < 6.00$ $19.00 < F_7 \leq 21.50$	$6.00 \leq F_4 < 8.25$ $16.90 < F_5 \leq 19.00$	$8.25 \leq F_2 < 10.25$ $14.00 < F_3 \leq 16.90$	$10.25 \leq F_1 \leq 14.00$	12.25
FLAG #2	Destruction level (P_6 and P_7)	Design Specification/ Performance tests (P_4 and P_5)	Robustness margin (P_2 and P_3)	Safety and approval tests (P_1)	Optimal operating point (P_0)
[bar]	$0.00 \leq P_6 < 0.38$ $2.12 < P_7 \leq 2.49$	$0.38 \leq P_4 < 0.63$ $1.81 < P_5 \leq 2.12$	$0.63 \leq P_2 < 1.00$ $1.50 < P_3 \leq 1.81$	$1.00 \leq P_1 \leq 1.50$	1.25
FLAG #3	Destruction level (T_6 and T_7)	Design Specification/ Performance tests (T_4 and T_5)	Robustness margin (T_2 and T_3)	Safety and approval tests (T_1)	Optimal operating point (T_0)
[°C]	$0.00 \leq P_2 < 10.00$ $75.00 < P_3 \leq 87.50$	$10.00 \leq T_4 < 22.50$ $65.00 < T_5 \leq 75.00$	$22.50 \leq T_2 < 31.25$ $55.00 < T_3 \leq 65.00$	$31.25 \leq T_1 \leq 55.00$	42.50

Therefore, information contained in the structures of the flags are used to evaluate the range of the feedback variables used in the control of the agricultural machines to support the decision making for realizing an accurate and adequate operation. In the same manner, the concept of the destruction level is related to the region wherein one can identify risks to the machinery's lifetime that must be avoided.

For the acquisition of a reference curve for the flow sensor, an ORION electromagnetic flowmeter, model Orion 4621A300000, installed at the outlet of the water pump of the ASDS was used [59]. The electromagnetic flowmeter had a measuring range of 5–100 l/min for pressures up to 4000 kPa. The calibration constant of this flowmeter, according to the manufacturer, was 600 pulses per liter, and the flow rate in liters per minute was obtained from a reading at a related frequency in Hertz. With the aid of the Arduino and the developed software, a group of reference flows in liters per minute was sent to the sensor, and a set of values was obtained from the sensor flow (Figure 8).

For the acquisition of a reference database with pressure values, a WIKA model A-10 pressure sensor was used. The voltage signals of the A-10 sensor varied from 0 to 10 V, proportional to their pressure measurement ranges from 0 to 16 bar, and this sensor had a reading error and a maximum linearity of 0.016 bar. With the aid of the Arduino and the developed software, considering intervals of 0.15 bar for a useful operating range of 0.5 to 3.0 bar, reference pressure values were sent to the pressure sensor, and the values obtained were recorded (Figure 9). In addition, to obtain a reference database with accurate temperature values, a calibrated sensor, type PT 100 of the Mit-Exact brand, was used, which was initially dipped in a beaker of water and ice. This water was heated with the aid of a mixer to 95 °C. As the temperature values increased, the internal resistance of the sensor also increased. For a better perception of the variation of the values of the sensor's resistance, a Wheatstone bridge was used. In this manner, it was possible to measure the unknown resistance of the sensor. The values were recorded at intervals of 5 °C, i.e., while taking into consideration an experimental range for the evaluation of different levels of the sprayer operation (Figure 10).

According to the flag structure for each variable, it is possible to perform, in real time, the agricultural sprayer's diagnosis, as well as, if actions are required, to find its prognostic and corrections based on the actuation by its control circuit, or even provide a recommendation for a sensor's replacement.

The prognostics and fault-tolerant strategies for reliable field operation can thus be obtained.

However, the transdisciplinary joint efforts of engineers and researchers are still required to fulfill the demands of such a field of knowledge and to promote the new paradigm shift in the reliability of agricultural machinery.

To illustrate the flexibility of the network used, Table III shows how the CAN messages were assembled.

TABLE III. CAN MESSAGES STRUCTURE AND TRANSMISSION INTERVAL

Sensor	CAN ID	DLC	Data (number of bits)			Transmission interval
Flow	100	3	LOW (8)	HI (8)	FLAG (2)	50ms
Pressure near flowmeter	101	3	LOW (8)	HI (2)	FLAG (2)	10ms
Pressure at boom	102	3	LOW (8)	HI (2)	FLAG (2)	10ms
Temperature	103	3	LOW (8)	HI (2)	FLAG (2)	10ms

As observed in Table III, the number of messages on the organized network is reduced. However, it is necessary to define a set of unique identifiers for each node. According to the resolution required for each variable, a number of bytes is allocated in the data frame, that is, as identified in the DLC field.

For example, the temperature and pressure variables are obtained via a 10-bit A/D converter, and, to maintain such a resolution, the data can be divided into two parts, LOW and HI, comprising 8 and 2 bits, respectively. However, the CAN controller does not transmit only the 2 bits that refer to the HI part of the data of interest. It transmits the total bytes, and this form of operation is indicated in the DLC field.

The transmissions of the flags along with the data facilitate the identification of the status of each sensor. In this context, it is important to note that, for each new node inserted, which implies a higher load on the bus, attention will be required to be focused to avoid instability in the CAN network, i.e., such a situation could result in a variable transmission time. Therefore, the WCTT must be kept in mind during the design of the control loop strategy in order to avoid destabilizing the control loop of a larger distributed architecture.

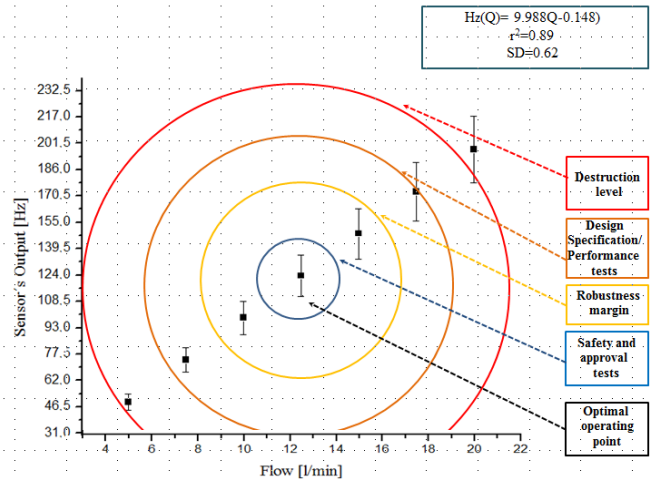


Figure 8. Reference curve for the flow sensor (electromagnetic flowmeter, model Orion 4621A300000) installed at the outlet of the water pump, and the experimental range results obtained for an agricultural sprayer's operation.

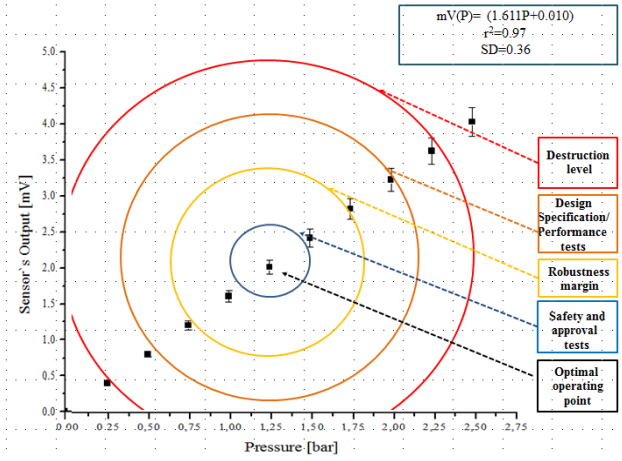


Figure 9. Reference curve for the pressure sensor (WIKA model A-10) installed at the boom, and the experimental range results obtained for an agricultural sprayer's operation.

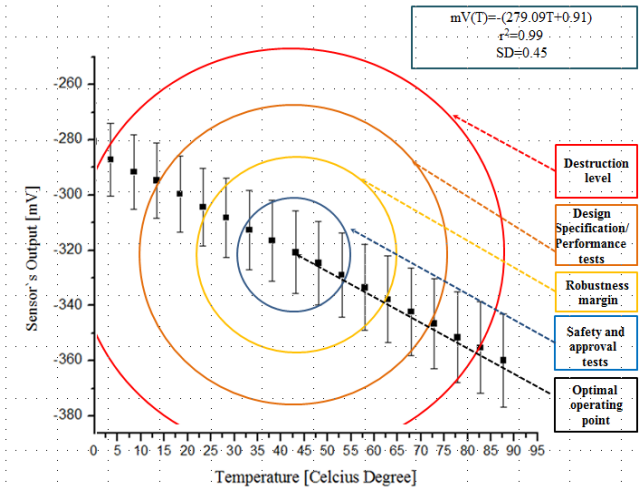


Figure 10. Reference curve for the temperature sensor (PT 100 of Mit-Exact) used to measure the temperature of the syrup, which is formed by adding the pesticides to water, and the experimental range results obtained for an agricultural sprayer's operation.

It is worth noting that, with the increase in the network complexity, it may be interesting to migrate to the latest version of CAN, known as CAN-FD (flexible data-rate), which allows a greater amount of data traffic than CAN 2.0, thus significantly reducing the load on the bus in terms of the utilization rate [60].

IV. CONCLUSIONS

An intelligent system for the evaluation of the failure and reliability of agricultural sprayers based on the sensors' information and a smart support decision-making architecture was presented. The obtained results showed that it is possible to observe real-time prognostics as well as to help with robustness to ensure quality aggregation in pest control processes based on agricultural spraying systems.

In addition, such a system enabled the configuration of a sensor's recalibration using an unsupervised algorithm while considering the use of a CAN bus protocol operating with the measurements of the flow rate, pressure, and temperature in the controlled circuit process of an agricultural sprayer. The proposed topology demonstrated feasibility for the implementation of the calibration modules, i.e., it benefited from the CAN networks, which are becoming widely used in agricultural machinery, based on the SAE J1939 standards collection. Furthermore, there are opportunities for the realization of real-time monitoring and fault-tolerant design hat can facilitate an extended lifetime and reduced failure rate as well as a better understanding of the failure mechanisms because more failure-mechanism-specific accelerated testing can be designed, which can result in improved reliability predictions for sensor-based agricultural machinery and its applications.

ACKNOWLEDGMENT

This research was supported by the Brazilian Corporation for Agricultural Research (Embrapa), Process No. 11.14.09.001.05.06, and the São Paulo Research Foundation (Fapesp), Process No. 17/19350-21.

REFERENCES

[1] P. E. Cruvinel, E. A. G. Penãloza, P. B. Andrade1, and H. V. Mercaldi, "Real-Time Failure and Reliability Analysis of Agricultural Sprayers Based on Sensors, Arduino Architecture, and Controller Area Bus Protocol", Proceedings of the IARIA, SENSORDEVICES 2019: The Tenth International Conference on Sensor Device Technologies and Applications, Nice, France, 2019, pp. 38-48.

[2] N. Wang, N. Zhang, and M. Wang "Wireless sensors in agriculture and food industry-recent development and future perspective", Science directs, Computers and Electronics in Agriculture, vol. 50, pp. 1-14, 2006..

[3] O. Elijah, T. A. Rahman, I. Orikumhi, C. Y. Leow and M. N. Hindia, "An Overview of Internet of Things (IoT) and Data Analytics in Agriculture: Benefits and Challenges", IEEE Internet of Things Journal, vol. 5, no. 5, pp. 3758-3773, 2018.

[4] Q. Li and C. Yao, Real-Time Concepts for Embedded Systems, CRC Press, pp. 287, 2017.

[5] J. Lin *et al*., "A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications", IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1125-1142, 2017.

[6] J. M. Antle, J. W. Jones, and C. E. Rosenzweig, "Next generation agricultural system data, models and knowledge products: introduction", Agricultural Systems, vol. 155, pp. 186-190, 2017.

[7] VDMA Verlag, Guideline Industrie 4.0, 2016. [retrieved: June, 2020]. Available from: https://www.vdma-verlag.com/home/artikel_72.html.

[8] L. Klerkx, E. Jakku, and P. Labarthe, "A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda", NJAS - Wageningen Journal of Life Sciences, vol. 90–91, pp. 16, 2019. [retrieved: June, 2020]. Available from: https://doi.org/10.1016/j.njas.2019.100315.

[9] D. C. Rose and J. Chilvers, "Agriculture 4.0: Broadenings Responsible Innovation in an era of smart", Farming Frontiers in Sustainable Food Systems, vol. 2, Issue 87, pp. 23, 2018. [retrieved: June, 2020]. Available from: https://doi.org/10.3389/fsufs.2018.00087.

[10] K. Pohl *et al*., Model-Based Engineering of Embedded Systems: The SPES 2020 Methodology, Springer Verlag, pp. 300, 2012.

[11] Food Security Information Network (FSIN): *Global Report on Food Crises (GRFC 2019)*, pp. 202, 2019. [retrieved: June, 2020]. Available from: http://www.fsinplatform .org/sites/default/files/resources/files/GRFC2019FullReport .pdf.

[12] Food and Agriculture Organization of the United Nations: "Global agriculture towards 2050", FAO, Rome, 2009.

[13] J. K. Schueller, "Automation and control". In: CIGR Handbook of Agricultural Engineering, Information Technology, vol. VI, ed. by A. Munack (CIGR, Tzukuba) pp. 184–195, Chap. 4, 2006.

[14] T. Stombaugh, E. Benson, and J. W. Hummel, "Automatic guidance of agricultural vehicles at high field speeds", ASAE Paper No. 983110 (ASAE, St. Joseph), 1998.

[15] M. Yitayew, K. Didan, and C. Reynolds, "Microcomputer based low-head gravity-flow bubbler irrigation system design", Computer and. Electronics in Agriculture, vol. 22, pp. 29–39, 1999.

[16] F. R. Miranda, R. E. Yoder, J. B. Wilkerson, and L. O. Odhiambo, "An autonomous controller for site-specific management of fixed irrigation systems, Computer and Electronics in Agriculture, vol. 468, pp. 183–197, 2005.

[17] S. Fountas *et al*., "Farm management information systems: current situation and future perspectives", Computer and. Electronics in Agriculture, vol. 115, pp. 40-50, 2015.

[18] C. G. Sørensen *et al*., "Conceptual model of a future farm management information system", Computer and Electronics in Agriculture, vol. 72, Issue 1, pp. 37-47, 2010.

[19] A. Kaloxylos *et al*., "Farm management systems and the future internet era", Computer and Electronics in Agriculture, vol. 89, pp. 130-144, 2012.

[20] M. Pecht and V. Ramappan, "Are components still the major problem: a review of electronic system and device field failure returns", IEEE Transaction on Components, Hybrids, Manufacturing Technology, vol, 15, pp. 1160-1164, 1992.

[21] C. T. Leonard, "Mechanical engineering issues and electronic equipment reliability: Incurred costs without compensating benefits", Journal of Electronic Packaging, vol. 113, pp. 1-7, 1991.

[22] A. Bar-Cohen, "Reliability physics vs reliability prediction", IEEE Transaction on Reliability, vol. 37, pp. 452, 1988.

[23] P. D. T. O'Connor, "Reliability prediction: Help or hoax?", Solid State Technology, vol. 33, pp. 59-61, 1990.

[24] H. V. Mercaldi, E. A. G. Peñaloza, R. A. Mariano, V. A. Oliveira, and P. E. Cruvinel, "Flow and Pressure Regulation for Agricultural Sprayers Using Solenoid Valves", In: International Federation of Automatic Control (IFAC), Elsevier, vol. 50, Issue 1, 2017, pp. 6607-6612, doi.org/10.1016/j.ifacol.2017.08.693.

[25] A. Marangoni Junior and M. C. Ferreira, "Influence of working pressure and spray nozzle on the distribution of spray liquid in manual backpack sprayers", Arquivos do Instituto Biológico, vol.86, 2019. [retrieved: June, 2020]. Available from: https://doi.org/10.1590/1808-1657000442018.

[26] T. Arvidsson, L. Bergström, and J. Kreuger, "Spray drift as influenced by meteorological and technical factors", Pest Management Science, vol. 67, Issue 5, pp. 586-598, 2011.

[27] ASAE S572. Spray nozzle classification by droplet spectra. In: ASAE Standards, St. Joseph, pp. 389-391, 2000.

[28] J. C. Christofoletti, Shell manual of machinery and techniques for the application of pesticides, (original in Portuguese Language: Manual Shell de máquinas e técnicas de aplicação de defensivos agrícolas), São Paulo, Shell, pp.124, 1992.

[29] D. L. Reichard, H. E. Ozkan, and R.D. Fox, "Nozzle wear rates and test procedure", Transaction of the ASAE, vol. 34, no. 6, pp. 2309-2316, 1991.

[30] ISO 16122-2, Agricultural and forestry machinery – Inspection of sprayers in use – Part 2: Horizontal boom sprayers, ISO Publication, pp. 18, 2015.

[31] H. Ganzelmeier and S. Rietz, "Inspection of plant protection in Europe". In: International Conference on Agricultural Engineering. Part II, Oslo. Proceedings, Eurageng, pp. 597-598, 1998.

[32] N. Bjugstad, Controlo of crop sprayers in Norway, Oslo: Ageng, Eurageng, 1998.

[33] M. A. Gandolfo, "Periodic inspection of agricultural sprayers" (original title: Inspeção periódica de pulverizados agrícolas). Doctoral Thesis, Faculty of Agronomic Sciences, São Paulo State University (UNESP), SP, Brazil, pp. 92, 2002.

[34] B. Huyghebaert *et al*., "Compulsory inspection of crop sprayers already in use Belgium", Selection of control method, Madrid: Ageng, CD-ROM, 1996.

[35] P. Braekman *et al*., "Organisation and results of the mandatory inspection of crop sprayers in Belgium", Belgium: Ministry of Small Enterprises, Traders and Agriculture, pp. 9, 2005.

[36] L. M. Val, "Programs for the formation of the applicators and equipment's maintenance" (original title: Programas de formación de aplicadores y programa de revisión de equipos. In: Jornada Internacional en Tecnologia de Aplicacion, 2007, Proceedings, Montevidéo: Republic University, CD-ROM, 2007.

[37] J. C. Magdalena and A. P. Di Prinzio, "Calibration service for fruit growing sprayers in Black River and Neuquén" (original title: Servicio de calibración de pulverizadoras frutícolas en Río Negro y Neuquén), In: Congreso Argentino de Ingenieria Rural, Córdoba. Proceedings.... Argentina: National Cordoba University, vol. 2, pp. 137-148, 1992.

[38] E. Fey, "State of the art in relation to the spray process of the associates from the COOPERVALE" (original title: Estado de arte do processo de pulverização junto a associados da COOPERVALE), Supervised Internship Report Ponta Grossa State University, Maripá – PR, Brazil, pp. 26, 1998.

[39] J. Langenakens and M. Pieters, "The organization and first results of the mandatory inspection of crop sprayers in Belgium", In: Aspects of applied Biology - Optimizing pesticide application, Agricultural Research Centre Ghent, pp. 233-240, 1997.

[40] S. Sartori, "Sprayers for tractor ground application" (original title: Pulverizadores para aplicação terrestre tratorizada), In: Simpósio Brasileiro sobre Tecnologia de

Aplicação de Defensivos Agrícolas, Jaboticabal, SP, Brazil, Proceedings... FUNEP, vol. 1, pp. 46-79, 1985.

[41] G. F. Dedordi *et al.*, "Technical-operational evaluation of bar sprayers in Pato Branco region" (original title: Avaliação técnica-operacional de pulverizadores de barras na região de Pato Branco), Acta Iguazu, Cascavel, PR, Brazil, vol. 3, no. 1, pp. 144-155, 2014.

[42] J. F. Schlosser, "Application technology and machine use: use of agrochemicals" (original title: Tecnologia de aplicação e uso de máquinas: uso de agroquímicos), Federal University of Santa Maria, Textbook - Technical Series, Module 5, 2002.

[43] F. Olivi and A. Simão, "In order to speed up the maintenance of sprayers, the diagnostic case provides technical support and spare parts" (original title: Com o objetivo de agilizar a manutenção dos pulverizadores, a maleta de diagnóstico dá suporte técnico e na reposição de peças). Agricultural News. [retrieved: February, 2020]. Available from: https://noticiasagricolas.com.br/videos/maquinas-e-tecnologias/.

[44] K. R. Felizardo, H. V. Mercaldi, P. E. Cruvinel, V. A. Oliveira, and B. L. Steward, "Modeling and model validation of a chemical injection sprayer system", Applied Engineering in Agriculture, vol. 32, Issue 3, pp. 285-297, 2016.

[45] X. Wei, S. Jian, and H. Sun, "Design and Test of Variable Rate Application Controller of Intermittent Spray Based on PWM", Journal of agricultural machinery vol. 43, Issue 12, pp. 87-129, 2012.

[46] H Liu, H Zhu, and Y Shen, "Development of digital flow control system for copy multi-channel variable-rate sprayers", Journal Transactions of the ASAE, vol. 57, Issue 1, pp. 273-281, 2014.

[47] American Society of Agricultural and Biological Engineers (ASAE) Standards: S572 - Spray nozzle classification by droplet spectra, no. 99, St. Joseph USA, pp. 389-391, 2000.

[48] P. E. Cruvinel *et al*., "An advanced sensors-based platform for the development of agricultural sprayers," In: Sergey Y. Yurish (Ed.). Sensors and Applications in Measuring and Automation Control Systems. [S. l.]: IFSA, pp. 181-204. (Advances in Sensors: Reviews; vol. 4), 2016.

[49] K. Etschberger, "Controller Area Network. Basics, Protocols, Chips and Applications", IXXAT, ISBN 3-00-007376-0, 2001.

[50] R. B. Gmb "CAN Specification, Version 2.0", Postfach 50, D-7000 Stuuugart 1, 1991. Imported into Framemaker 4 by Chuck Powers, BOSCH, Motorola MCTG Multiplex Applications, April 3, 1995.

[51] E. Godoy *et al*., "Design of CAN-based distributed control systems with optimized configuration", Journal of the Brazilian Society of Mechanical Sciences and Engineering, vol. 32, no. 4, pp. 420-426, 2010.

[52] B. Woonhyuk, J. Seyong, S. Hoin, K. Soontae, S. Bongsob, C. Dongkyoung, "A CAN-based Distributed Control System for Autonomous All-Terrain Vehicle (ATV)", 17th World Congress of The International Federation of Automatic Control, IFAC Proceedings, Seoul, Korea, vol. 41, Issue 2, pp. 9505-9510, 2008.

[53] J. R. Edwards, "Polynomial regression and response surface methodology," In: C. Ostroff & T. A. Judge (Eds.), Perspectives on organizational fit, pp. 361–372. San Francisco: Jossey-Bass, 2007.

[54] L. Tornqvist, P. Vartia, and Y. Vartia, "How should relative changes be measured?" The American Statistician, vol. 39, Issue 1, pp. 43–46, 1985.

[55] P. E. Danielsson, "Euclidean distance mapping," Computer Graphics Image Processing, vol. 14, pp. 227-248, 1980.

[56] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - arguments against avoiding RMSE in the literature," Geoscience Model Development, vol. 7, pp. 1247–1250, 2014.

[57] R. H. Walden, "Analog-to-digital converter survey and analysis", IEEE Journal on Selected Areas in Communication, vol. 17, no. 4, pp. 539–550, 1999.

[58] TJA 1050 Data Sheet – High speed CAN transceiver. Product specification, Integrated Circuits, Philips Semiconductors, pp. 18, 2003.

[59] ORION Electro-Magnetic Flow Meter. [retrieved: June, 2020]. Available from: mcpheeenterprises .com/orion-electromagnetic-flowmeter/.

[60] R. de Andrade *et al.*, "Analytical and Experimental Performance Evaluations of CAN-FD Bus", IEEE Access. 10pp., 2017. [retrieved: June, 2020]. Available from: file:///C:/Users/HP/Downloads/v38-CAN-FD-Performance-10APR2018-Max2_Publicada.pdf.

Fast FPGA-Placement Using a Gradient Descent Based Algorithm

Timm Bostelmann, Tobias Thiemann and Sergei Sawitzki

FH Wedel (University of Applied Sciences)
Wedel, Germany
Email: {bos,inf103917,saw}@fh-wedel.de

***Abstract*—Programmable circuits and, nowadays, especially Field-Programmable Gate Arrays (FPGAs) are widely applied in computationally demanding signal processing applications. Considering modern, agile hardware / software codesign approaches, an Electronic Design Automation (EDA) process not only needs to deliver high quality results, but also has to be swift because software compilation is already distinctly faster. Slow EDA tools can in fact act as a kind of show-stopper for an agile development process. One of the major problems in EDA is the placement of the technology-mapped netlist to the target architecture. In this work, a method to reduce the runtime of the netlist placement for FPGAs is evaluated. The approach is a variation of analytical placement, with the distinction that a gradient descent is used for the optimization of the placement. This work is an extended version of a previous publication of the authors on this topic. Additionally, it is based on previous publications of the authors, in which a placement algorithm using self-organizing maps is introduced and optimized. In comparison, the gradient placement approach is shown to be up to 3.8 times faster than the simulated annealing based reference. The quality regarding the critical path is shown to be about 43 percent worse on average. The bounding-box and routing-resource costs are shown to be about equal to the reference.**

Keywords–EDA; FPGA; placement; gradient descent.

I. Introduction

The ever-growing complexity of Field-Programmable Gate Arrays (FPGAs) has a high impact on the performance of Electronic Design Automation (EDA) tools. A complete compilation from a hardware description language to a bitstream can take several hours. One step highly affected by the vast size of netlists is the NP-equivalent placement process. It consists of selecting a resource cell (position) on the FPGA for every cell of the applications netlist. In this work, a fast placement algorithm is benchmarked and evaluated. It is an extended version of [1] with a detailed visual representation of the original data and additional benchmarks based on a timing analysis. In previous publications of the authors, a placement algorithm for FPGAs based on a self-organizing map [2] was presented [3] and optimized [4]. With that approach, placements of high quality were produced. However, it was relatively slow for large netlists, even when accelerated using a Graphics Processing Unit (GPU) [5]. Therefore, in this work, a faster approach for netlist placement based on a gradient descent is presented as an updated version of the authors' previous work [5].

Due to the complexity of the netlist placement problem, many current algorithms work in an iterative manner. A well known example is simulated annealing [6], which starts with a random initial placement and swaps blocks stepwise. The result of every step is evaluated by a cost-function. A step is always accepted, if it reduces the cost. If it increases the cost, it is accepted with a probability that declines with time (cooling down). An annealing schedule determines the gradual decrease of the temperature, where a low temperature means a low acceptance rate and a high temperature means a high acceptance rate. Generally, the temperature T_n is described by an exponentially falling function like

$$T_n = \alpha^n \cdot T_0, \tag{1}$$

where T_0 is an initial temperature and α is an empirical constant, typically chosen as $0.7 \leq \alpha \leq 0.95$. However, there has been a lot of research on the optimization of the annealing schedule like in [7][8]. As a result, there are many variations available for any related problem.

Analytical placement is a different approach, where the problem is described as a system of equations. By solving this system of equations, the optimal position for every element can be derived. However, solving such large equation systems takes much time. Therefore, Vansteenkiste et al. [9] have introduced a method to approximate the solution of the equation system by the steepest gradient descent. This approach is shown to be two times faster than a conventional analytical placement on average, without any penalties in quality.

In this work, a simplified implementation of the steepest gradient descent placement is described and benchmarked extensively. It is not compared to other analytical placement methods. Instead, the established implementation of the simulated annealing approach of the Versatile Place and Route (VPR) tool [10] for FPGAs is used as reference.

In Section II, the problem of netlist placement for FPGAs is introduced and the principle of netlist placement with a gradient descent is described. In Section III, the proposed algorithm is described including some details of its implementation. In Section IV, the results of the proposed algorithm are presented. As representation for real world applications, a set of twenty Microelectronics Center of North Carolina (MCNC) benchmarks [11] is used. Based on these benchmarks, the bounding-box costs, the necessary channel width, the total wire length, the runtime, the total net delay and the critical path delay are analyzed and evaluated. Finally, in Section V, the results of this work are summarized and a prospect to further work is given.

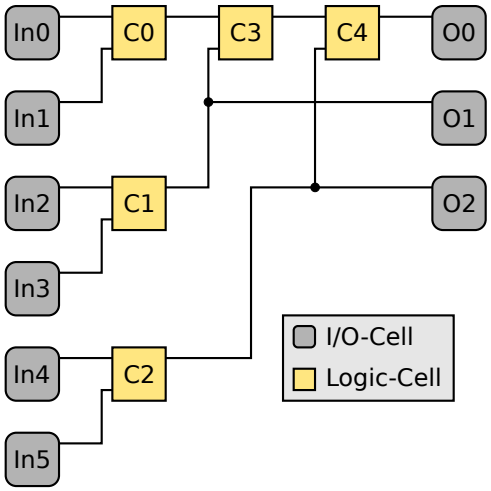


Figure 1. An exemplary graph of a netlist consisting of input-, output-, and logic-cells.

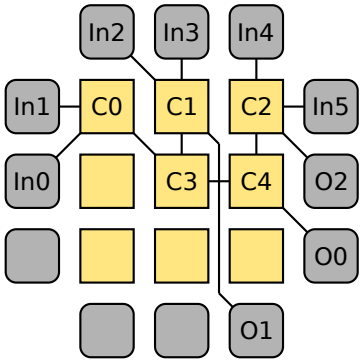


Figure 2. A valid, good placement for the previously introduced exemplary graph of a netlist on a simple island-style FPGA architecture.

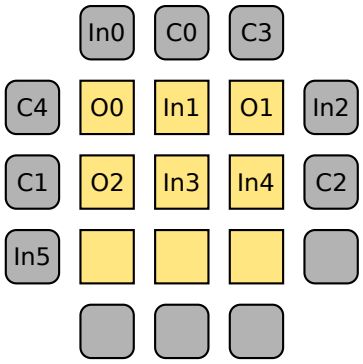


Figure 3. An invalid / illegal placement for the previously introduced exemplary graph of a netlist on a simple island-style FPGA architecture with usage of incompatible cell types.

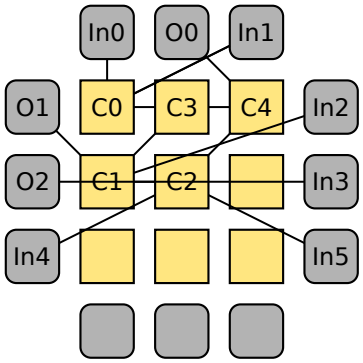


Figure 4. A valid, unfavorable placement for the previously introduced exemplary graph of a netlist on a simple island-style FPGA architecture.

II. Background

This section is separated into two parts. First, the problem of netlist placement for FPGAs is introduced. Second, the general idea of using a gradient descent for the placement of netlists for FPGAs is described.

A. Netlist Placement for FPGAs

The problem of netlist placement for FPGAs can be roughly described as selecting a resource cell (a position) on the target FPGA for every cell of the given netlist. In Figure 1, an exemplary graph of a netlist is defined. An exemplary placement for this netlist is presented in Figure 2. The positions must be chosen in a way that:

1) Every cell of the netlist is assigned to a resource cell of the fitting type (e.g., Input / Output or Logic).
2) No resource cell is occupied by more than one cell of the netlist.
3) The cells are arranged in a way that allows the best possible routing.

The first two rules are necessary constraints. A placement that is failing at least one of these two constraints is illegal and, therefore, unusable. For example, the placement shown in Figure 3 is illegal, because it fails the first constraint. Specifically, several logic-cells (e.g., C0 and C1) are placed on Input / Output-positions and several Input / Output-cells (e.g., In1 and In3) are placed on logic positions. The third rule is a quality constraint, which is typically described by a cost-function. The goal of a placement algorithm is to optimize the placement regarding this function without violating one of the necessary constraints. Usually, the length of the critical path and the routability are covered by the cost-function. For example, the placements shown in Figure 4 and Figure 2 are both valid. However, Figure 4 will clearly result in a longer critical path and a higher routing channel load.

B. Netlist Placement With a Gradient Descent

The netlist placement with a gradient descent is done by iteratively optimizing the positions of all elements of the netlist in the direction of the steepest gradient descent. During this process, the nodes are not bound to the grid of the FPGA architecture. Instead, they are positioned in a continuous space. To generate a valid placement – without overlapping and under consideration of the FPGA's architecture – in this approach, a cycle of optimization and legalization is used. This procedure is customary for analytical placement methods for FPGAs, like Gort and Anderson have introduced in [12]. A different approach would be to generate only valid placements by exclusively moving the nodes on the architectural grid of the FPGA.

III. Implementation

This section is separated into four parts, describing the implementation of the gradient calculation, the legalization,

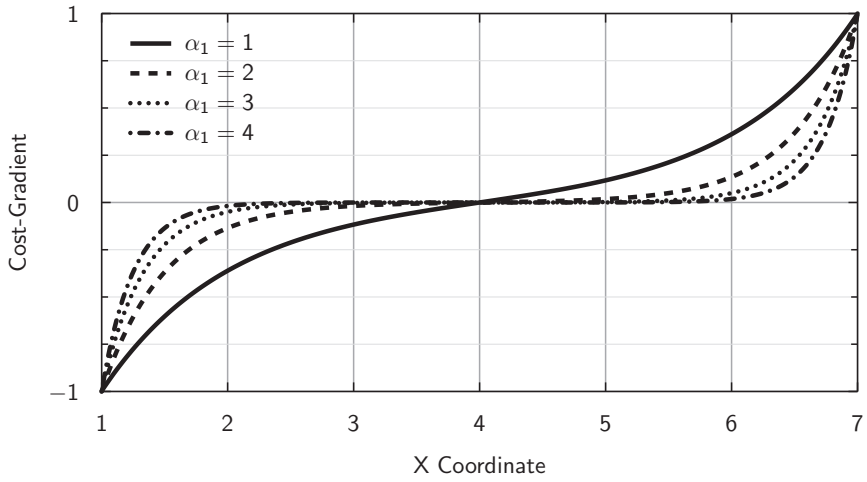


Figure 5. Exemplary plot of possible gradients for the X coordinate of a node, assuming a net with the boundaries $\min_x = 1$ and $\max_x = 7$.

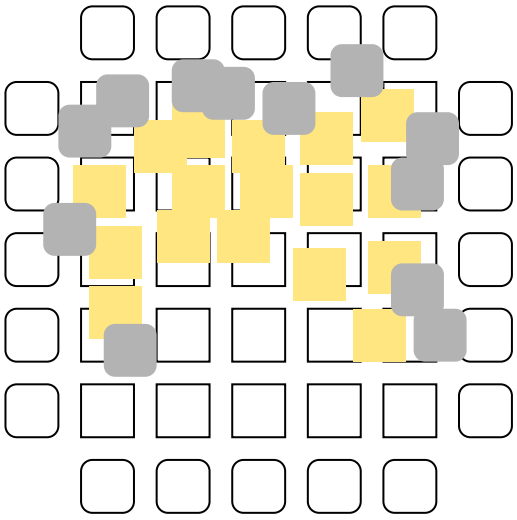


Figure 6. Exemplary placement before the legalization step.

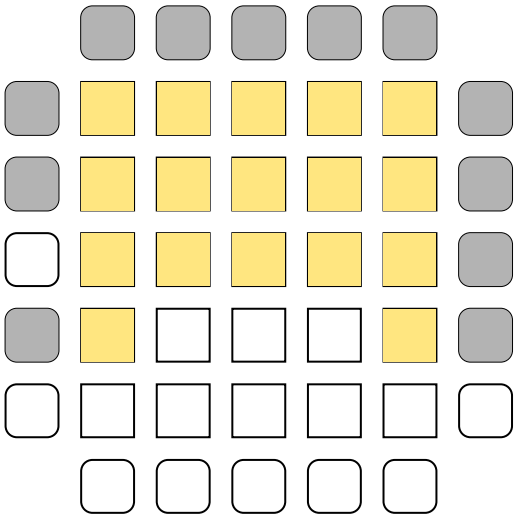


Figure 7. Exemplary placement after the legalization step.

the optimization and the placement phases.

A. Gradient Calculation

At the beginning of every optimization step, the bounding-box size of every net in the netlist is determined. This is a necessary preparation for the cost-function, which is described later in this section. To determine the size of a net, all nodes with a connection to the net are determined. For all these nodes, the minimum and maximum of the horizontal positions (X_i) and the vertical positions (Y_i) are determined and stored for the calculation of the gradient. Additionally, the sum of all sizes in X and Y direction is calculated, as a metric for the global quality of the current placement.

The goal of every optimization step is to move the nodes in a direction that leads to a reduction of the bounding-box size of the containing net. A cost-function is necessary to determine the influence of every node on the size of the corresponding net. The gradient of this cost-function can then be used to determine the direction of the movement of each node. All nodes of the netlist are moved towards the steepest gradient descent to reduce the global cost.

An intuitive approach would be to use the sum of the bounding-box sizes of all nets as cost-function. However, with this metric, only the outermost nodes would be moved and even nodes that are very near to the bounding-box would be ignored. Furthermore, the min and max functions contained in the metric can not be derived to calculate the gradient.

To solve these issues, an exponential function over the distance between the position of the node and the bounding-box of the net is chosen as basis of the cost-function. The cost-function for a node with the index k is

$$C_k = \alpha_2 \cdot \sum_{n \in N_k} \left(e^{\alpha_1 \cdot (x_k - \max_x(n))} + e^{\alpha_1 \cdot (\min_x(n) - x_k)} + e^{\alpha_1 \cdot (y_k - \max_y(n))} + e^{\alpha_1 \cdot (\min_y(n) - y_k)} \right), \quad (2)$$

where x_k and y_k describe the X and Y coordinates of the current node, N_k describes the set of all nets that contain the node and $\min_x$, $\max_x$, $\min_y$ and $\max_y$ are the minimal and maximal coordinates of the current net (i.e., the bounding-box). α_1 and α_2 are parameters for the cost-function, which

allow to influence the behavior of the function. α_1 determines how large the distance between the node and the bounding-box must be to reduce its influence in the cost-function. α_2 increases or reduces the cost to influence the steepness of the gradient. The influence of α_1 on the gradient is shown in Figure 5 for the X coordinate of a node, assuming a net with the boundaries $\min_x = 1$ and $\max_x = 7$. These boundaries were picked exemplarily to show the effect of the distance between the node and the bounding-box on the cost-function. It can be seen that at

$$X = \frac{\min_x + \max_x}{2} = \frac{1+7}{2} = 4, \tag{3}$$

in the center of the bounding-box, the cost is zero. In both directions, the absolute value of the cost is rising up to a limit of ± 1 at $\min_x$ and $\max_x$, effectively pushing the node to the center of the bounding-box.

The gradients for the X and Y coordinates can be calculated as the partial derivatives of (2) as in

$$\frac{\partial C_k}{\partial x_k} = \alpha_2 \cdot \sum_{n \in N_k} \left(e^{\alpha_1 \cdot (x_k - \max_x(n))} - e^{\alpha_1 \cdot (\min_x(n) - x_k)} \right), \tag{4}$$

$$\frac{\partial C_k}{\partial y_k} = \alpha_2 \cdot \sum_{n \in N_k} \left(e^{\alpha_1 \cdot (y_k - \max_y(n))} - e^{\alpha_1 \cdot (\min_y(n) - y_k)} \right). \tag{5}$$

As a result, the coordinates of nodes that are near the bounding-box of their containing net have a gradient of $\pm\alpha_2$, where the coordinates of nodes with a larger distance to the bounding-box have a much lower gradient, as shown in Figure 5. Consequentially, nodes with a larger gradient value must be moved further to optimize the placement quality.

B. Legalization

During the optimization step, the nodes can take any position. Thereby, illegal placements are produced, due to overlapping of nodes, as well as violation of the architectural grid of the FPGA. Therefore, the optimized placement must be legalized in a separate step. This is done by finding the nearest valid position for every node, as depicted in Figure 6 (before the legalization) and Figure 7 (after the legalization).

The algorithm for the legalization is inspired by the work of Gort and Anderson [12]. The basic idea of that approach is to find regions that contain more nodes than the corresponding region of the FPGA provides. Then, those regions are gradually expanded. When two regions overlap, they are merged. This is done until the regions are large enough to place all contained nodes to a proper resource cell of the FPGA. In the next step, the regions are split recursively and the nodes are assigned to the new sections by their position. This is repeated until a region contains no more nodes, or only one node. In the latter case, the position of the single remaining node is set to the position of its containing region.

In this work, the search for regions that contain more nodes than the corresponding region of the FPGA provides and the following expansion and merge phases are skipped. Instead, all nodes are assigned to one large region from the start and the phase of recursive splitting starts directly. By this measure, the computational effort for the legalization is reduced significantly without a dramatic impact on the global quality. This is because – especially when a large amount of the available resources is used – the result of the expansion phase is containing usually very few large regions or often only one large region anyway.

C. Optimization

For the optimization, the algorithm Adam – which was introduced by Kingma and Ba in [13] – is used. The used update rules are:

$g_t = \Delta\phi_t$	Gradient of the variable
$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t$	Running average force one
$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2$	Running average force two
$\hat{m}_t = m_t / (1-\beta_1^t)$	Bias corrected force one
$\hat{v}_t = v_t / (1-\beta_2^t)$	Bias corrected force two
$\phi_t = \phi_{t-1} - S_a \cdot \hat{m}_t / \left(\sqrt{\hat{v}_t} + \epsilon\right)$	Update of the variable

The constants β_1 and β_2 define how fast the averages of the first and second forces change. In this work, the constants were defined as $\beta_1 = 0.96$ and $\beta_2 = 0.998$. The variable S_a defines the learning rate or, more specifically, the step-width. It starts at $S_a = 1.5$, but changes over time (i.e., in the different phases of the placement). The Adam algorithm and the meaning of its update rules are described in full detail in the original source [13, Section 2].

D. Placement Phases

The previously described steps are executed for every iteration. The placement process is separated into five phases, with different parameters. Each phase consists of a given number of iterations. The number of iterations per phase was determined empirically and is fixed (i.e., independent of the size of the design). The phases are:

1) Presorting (5000 iterations)
   In this phase, all nodes are moved with a high step width in the general direction of their final position.
2) Grid placement (1000 iterations)
   In this phase, the force of the legalization is increased. Thereby, the nodes are pulled harder towards legal positions (i.e., to fitting cells of the architecture). This is necessary – for example – to prevent input and output cells from getting stuck in the logic block section of the architecture.
3) Initial detailed placement (1000 iterations)
   In this phase, the global step-width is reduced to one tenth of the initial value. This influences the legalization and the optimization equally, so that the balance between those two steps is not changed. However, the changes are much smaller, resulting in a more precise outcome.
4) Detailed placement (5000 iterations)
   In this phase, the step-width of the optimization is reduced linearly to 20 percent of its original value. Thereby, the nodes are pulled relatively harder towards their final positions in the grid.
5) Final placement (100 iterations)
   In this phase the influence of the optimization is reduced to zero, so that effectively only the legalization is active. Hence, the nodes are moved to their final position in the grid.

IV. Results

In this section, the benchmark results of the previously described placement algorithm are presented. VPR is used as reference for the comparison of the placement results, as well as for the routing and timing analysis.

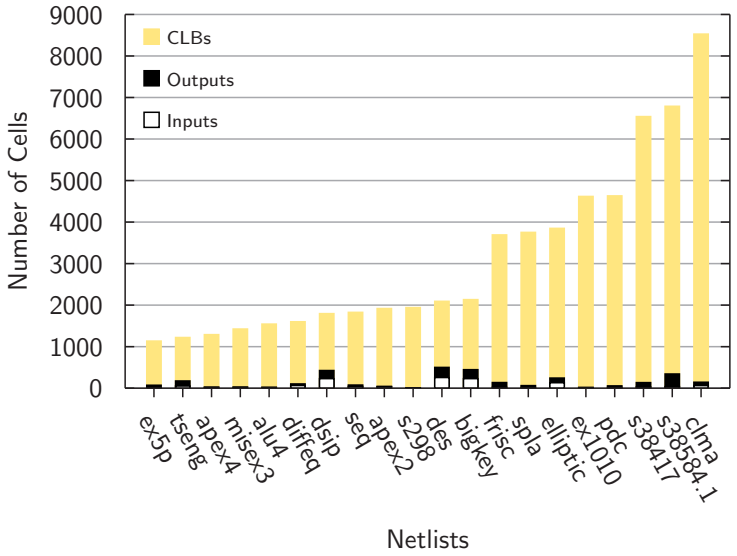


Figure 8. Diagram of the used benchmarks and their characteristics, the number of CLBs, input blocks, output blocks, sorted by the global block count.

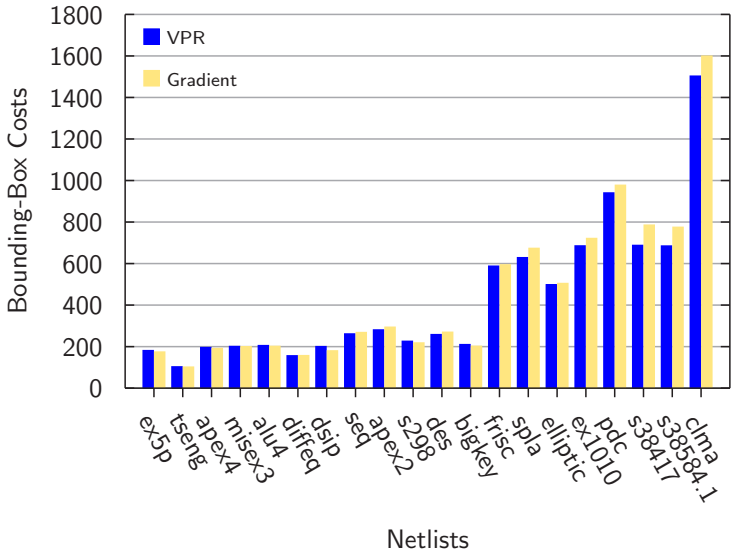


Figure 9. Diagram of the bounding-box costs of the gradient placement and the simulated annealing of VPR.

All used MCNC benchmarks [11] and their characteristics, namely, the number of Configurable Logic Blocks (CLBs), input blocks, output blocks and the sum of all blocks are listed in Table I, sorted by ascending complexity (i.e., the global block count). Additionally, in Figure 8, a detailed visual representation of the data is provided to underline the different distributions of block types between the netlists. The netlists are placed on a homogeneous island-style architecture with four input lookup tables.

A. *Bounding-Box Costs*

The standard metric used for the approximation of the quality of a placement in VPR is the bounding-box cost. It is basically the sum of the half perimeter of the bounding-boxes (i.e., length plus width) of all nets. As introduced by Betz and Rose in [10], the bounding-box metric can be described as

$$Cost = \sum_{n=1}^{N_{nets}} q(n) \cdot \left(\frac{bb_x(n)}{C_{av,x}(n)} + \frac{bb_y(n)}{C_{av,y}(n)} \right), \quad (6)$$

where $bb_x(n)$ and $bb_y(n)$ describe the horizontal and vertical size of the net n. $C_{av,x}(n)$ and $C_{av,y}(n)$ describe the average capacity of horizontal and vertical channels in the region of the net (in the considered case, the capacity is homogeneous over the whole architecture, so these values are constant. $q(n)$ corrects the effort for nets with more than three terminals, because it would otherwise be approximated to low.

In Table II, the bounding-box costs for the previously introduced benchmark netlists are presented. The results of VPR and the gradient placer are shown as absolute values and in relation to each other:

$$Cost_{Relative} = \frac{Cost_{VPR}}{Cost_{Gradient}} \cdot 100\,\% \quad (7)$$

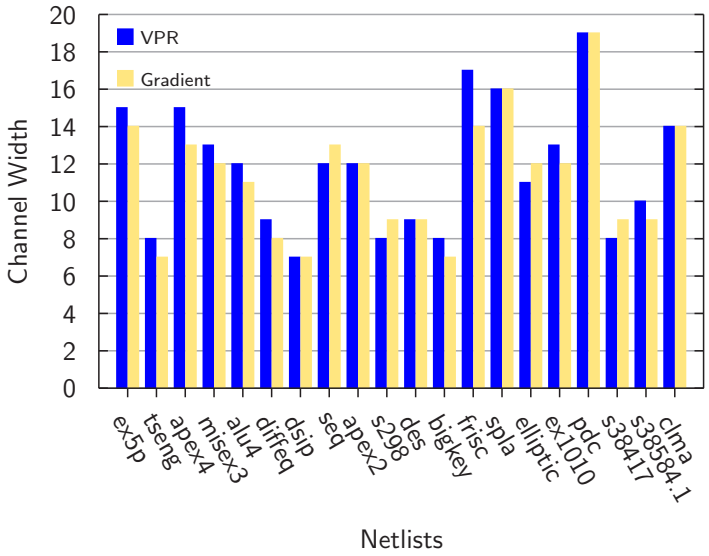


Figure 10. Diagram of the channel width of the gradient placement and the simulated annealing of VPR.

TABLE I. A LIST OF THE USED BENCHMARKS AND THEIR CHARACTERISTICS, THE NUMBER OF CLBS, INPUT BLOCKS, OUTPUT BLOCKS AND THE GLOBAL BLOCK COUNT

Name	Inputs	Outputs	CLBs	Blocks
ex5p	8	63	1064	1135
tseng	52	122	1047	1221
apex4	9	19	1262	1290
misex3	14	14	1397	1425
alu4	14	8	1522	1544
diffeq	64	39	1497	1600
dsip	229	197	1370	1796
seq	41	35	1750	1826
apex2	38	3	1878	1919
s298	4	6	1931	1941
des	256	245	1591	2092
bigkey	229	197	1707	2133
frisc	20	116	3556	3692
spla	16	46	3690	3752
elliptic	131	114	3604	3849
ex1010	10	10	4598	4618
pdc	16	40	4575	4631
s38417	29	106	6406	6541
s38584.1	38	304	6447	6789
clma	62	82	8383	8527

TABLE II. COMPARISON OF THE BOUNDING-BOX COSTS BETWEEN THE GRADIENT PLACEMENT AND THE SIMULATED ANNEALING OF VPR

Netlist	VPR	Gradient	Relative / %
ex5p	180.599	173.701	96.18
tseng	102.398	101.112	98.74
apex4	195.338	190.657	97.60
misex3	200.456	199.160	99.35
alu4	204.692	200.965	98.18
diffeq	155.531	156.375	100.54
dsip	199.845	179.254	89.70
seq	260.789	267.686	102.64
apex2	280.120	293.168	104.66
s298	225.344	217.479	96.51
des	257.643	268.889	104.36
bigkey	209.470	201.344	96.12
frisc	587.227	593.630	101.09
spla	628.155	672.990	107.14
elliptic	497.645	503.854	101.25
ex1010	684.798	720.589	105.23
pdc	939.813	976.890	103.95
s38417	687.198	784.862	114.21
s38584.1	684.220	774.451	113.19
clma	1502.330	1598.670	106.41
Average			101.85

Additionally, in Figure 9, a bar-graph of the data is provided to underline the different effects on the netlists. It can be seen that especially the smaller netlists profit from the gradient placement. Remarkably, for all netlists with less than 1600 nodes, the bounding-box costs are less with the gradient placer than with VPR. If the larger netlists are included, the costs for the gradient placer are only 1.85 percent higher on average, which is almost equal.

B. Channel Width and Wire Length

After their generation, the placements were routed with the VPR router and the Channel Width (CW), as well as the amount of necessary wire elements as a measure for the total Wire Length (WL) were determined. The results are shown in Table III. The differences in the channel width are given as a simple difference between the results:

$$\Delta CW = CW_{VPR} - CW_{Gradient} \tag{8}$$

The differences in the wire length are given as ratio between the results in percent:

$$WL_{Relative} = \frac{WL_{VPR}}{WL_{Gradient}} \cdot 100\,\% \tag{9}$$

Additionally, in Figure 10 and Figure 11, bar-graphs of the data are provided to underline the different effects on the netlists. The needed channel width of the gradient method is on average

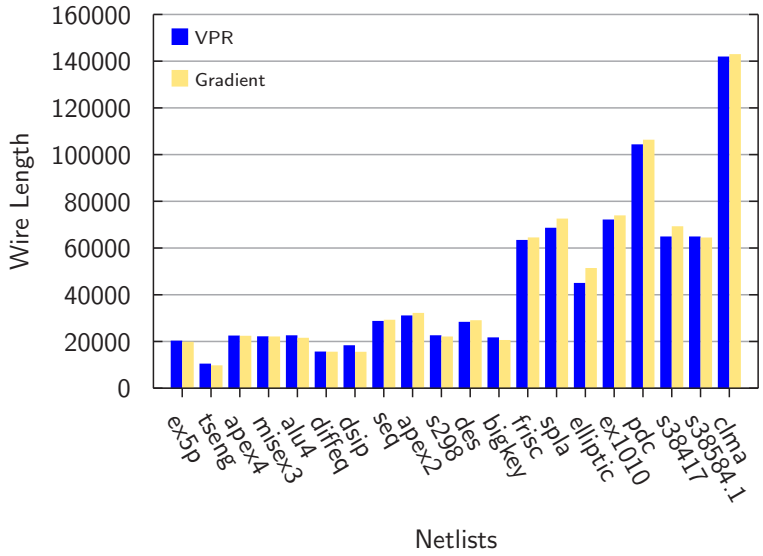


Figure 11. Diagram of the wire length of the gradient placement and the simulated annealing of VPR.

TABLE III. COMPARISON OF THE MINIMAL CHANNEL WIDTH (CW) AND THE TOTAL WIRE LENGTH (WL) BETWEEN THE GRADIENT BASED PLACEMENT ALGORITHM AND THE SIMULATED ANNEALING OF VPR

Netlist	VPR		Gradient		Relative	
	CW	WL	CW	WL	ΔCW	WL / %
ex5p	15	20034	14	19541	-1	97.54
tseng	8	10200	7	9463	-1	92.77
apex4	15	22215	13	22116	-2	99.55
misex3	13	21884	12	21820	-1	99.71
alu4	12	22319	11	21261	-1	95.26
diffeq	9	15369	8	15292	-1	99.50
dsip	7	18065	7	15260	0	84.47
seq	12	28469	13	28977	1	101.78
apex2	12	30826	12	31905	0	103.50
s298	8	22335	9	21801	1	97.61
des	9	28084	9	28764	0	102.42
bigkey	8	21424	7	20315	-1	94.82
frisc	17	63146	14	64220	-3	101.70
spla	16	68364	16	72288	0	105.74
elliptic	11	44742	12	51127	1	114.27
ex1010	13	71891	12	73653	-1	102.45
pdc	19	104065	19	106057	0	101.91
s38417	8	64626	9	68999	1	106.77
s38584.1	10	64626	9	64180	-1	99.31
clma	14	141660	14	142695	0	100.73
Average					-0.5	100.09

0.5 channels smaller than the reference, whereas its total wire length is 0.09 percent longer. Both values are considered to be almost equal to the reference.

C. Runtime

In the previous sections, it was shown that the gradient placer produces a similar placement quality as VPR in regard to the bounding-box cost and the required routing resources. In this section, the runtime of both algorithms is measured and evaluated. The configuration of the system that has been used for the benchmarking is provided in Table IV.

The results are shown in Table V. The presented numbers

TABLE IV. CONFIGURATION OF THE SYSTEM THAT HAS BEEN USED FOR THE BENCHMARKING OF THE GRADIENT ALGORITHM AND VPR

Property	Value
Processor	Intel® Core™ i7-4510U
Cores	2
Threads	4
Base Frequency	2.00 GHz
Turbo Frequency	3.10 GHz
Cache	4 MB
RAM	16 GB

are each an average of ten measurements. All single measurements varied less than two percent of the average of the measurement series.

On average, the gradient based placement algorithm needs less than half of the time of the simulated annealing placer of VPR. Furthermore, the ratio is even better for large netlists, as can be seen clearly in Figure 12. For example, the largest netlist in this benchmark series – the clma netlist – is placed 3.8 times faster with the gradient based approach.

D. Delay

In the conclusion of [1], a longer critical path was mentioned as a possible drawback of the gradient placer in comparison with VPR. However, this was only a generalized conclusion based on preliminary results. Therefore, in this work, detailed comparisons of the resulting Total Net Delay (TND) and Critical Path Delay (CPD) are presented. The maximum clock speed of the resulting implementation is determined by the CPD. Therefore, the CPD is a good metric for the speed a circuit can run at. In Table VI, the results of CPD comparison are presented. Additionally, in Figure 13, a bar-graph of the data is provided to underline the different effects on the netlists. As expected by the preliminary results, the CPD of the gradient based placement is consistently worse than the reference. The CPD of the gradient based placement algorithm is about 43 percent longer than the CPD of the reference. This is due to the fact, that the cost-function of

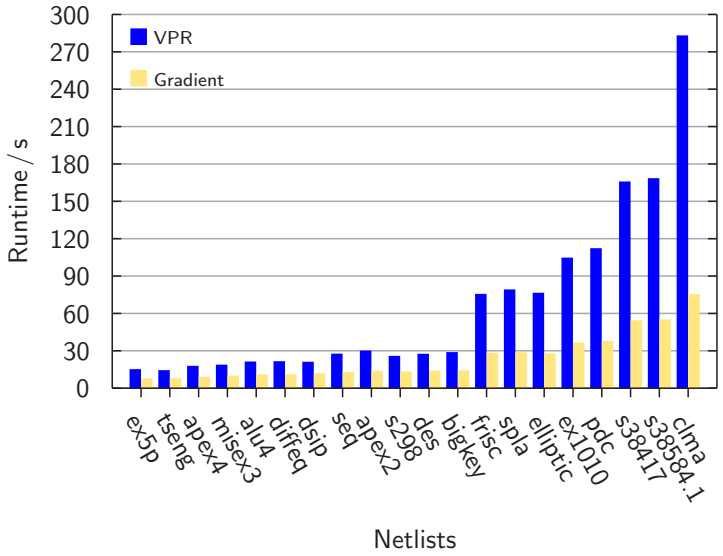


Figure 12. Diagram of the runtime as average of ten measurements between the gradient based placement algorithm and the simulated annealing of VPR.

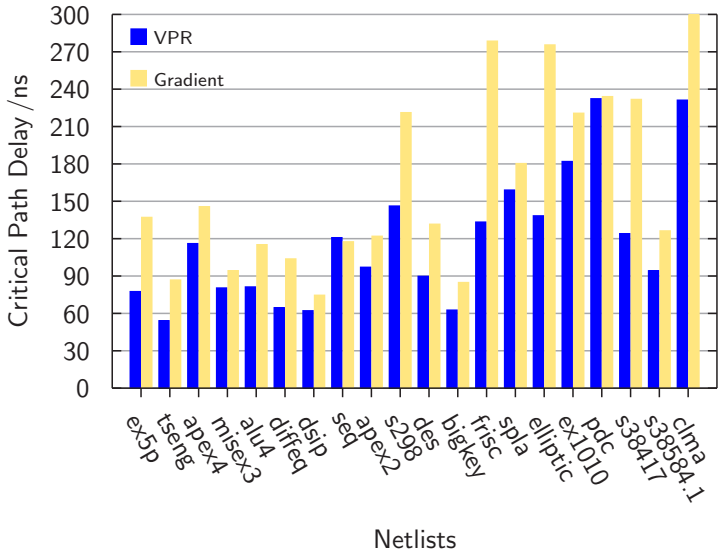


Figure 13. Diagram of the Critical Path Delay (CPD) of the gradient placement and the simulated annealing of VPR.

VPR optimizes the critical path directly, while the presented gradient based placement algorithm only optimizes the global delay without distinction of the critical path.

The TND represents the total delay of all routing structures in the critical path (i.e., without the delay of the CLBs). In Table VII, the results of TND are presented. Additionally, in Figure 14, a bar-graph of the data is provided to underline the different effects on the netlists. Similar to the CPD, the TND of the gradient based placement algorithm is about 46 percent longer than the TND of reference.

The higher delays (i.e., CPD and TND) are not an intrinsic problem of the gradient based placement algorithm. They are rather a result of the cost-function, which is – in its current form – not tailored towards CPD optimization. Preliminary results show that a modified cost-function will indeed improve the CPD and TND in exchange for slightly worse bounding-box costs. The basic concept is to determine the length of each path in the netlist and increase or decrease the gradient's steepness based on the criticality of the corresponding node.

V. Conclusion and Future Work

In this work, a fast approach for netlist placement based on a gradient descent was presented. The gradient placer was compared to the simulated annealing based placer of VPR. It has been shown that the quality of the placement in regard of the bounding-box cost and the occupation of routing resources (i.e., channel width and total wire length) is equal to the reference within a reasonable margin of error, as proven by placing twenty prominent benchmarking netlists of different complexity. Notably, the presented approach is shown to be up to 3.8 times faster than the reference. On average, it needs less than half of the time to compute the result. However, it also has been shown that the resulting length of the critical path is worse with the gradient placer (about 43 percent on average). A possible solution to this issue (i.e., a modified cost-function)

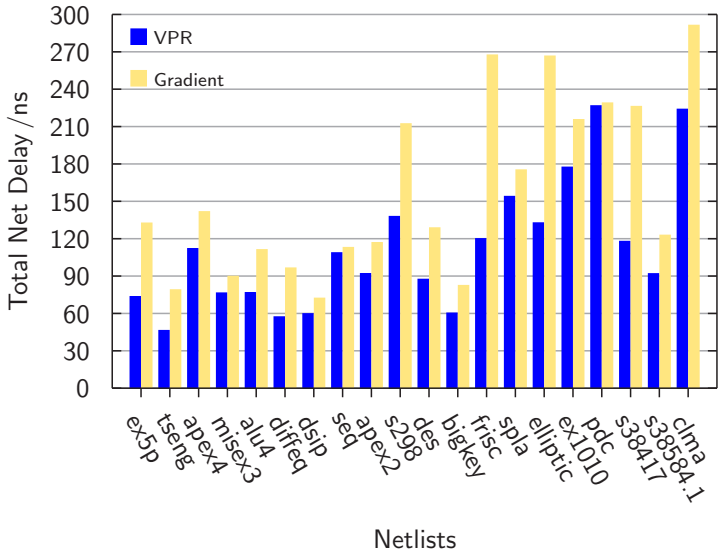


Figure 14. Diagram of the Total Net Delay (TND) of the gradient placement and the simulated annealing of VPR.

TABLE V. COMPARISON OF THE RUNTIME AS AVERAGE OF TEN MEASUREMENTS BETWEEN THE GRADIENT BASED PLACEMENT ALGORITHM AND THE SIMULATED ANNEALING OF VPR

Netlist	VPR / s	Gradient / s	Relative / %
ex5p	14.69	7.23	49.23
tseng	13.86	7.34	53.00
apex4	17.34	8.53	49.16
misex3	18.27	9.54	52.20
alu4	20.81	10.48	50.36
diffeq	21.05	10.63	50.47
dsip	20.62	11.45	55.54
seq	27.12	12.53	46.21
apex2	29.60	13.41	45.31
s298	25.35	12.90	50.90
des	27.01	13.48	49.92
bigkey	28.36	13.72	48.36
frisc	75.10	27.83	37.05
spla	78.67	28.21	35.85
elliptic	76.02	27.79	36.56
ex1010	104.21	36.11	34.65
pdc	111.76	37.30	33.37
s38417	165.32	53.89	32.60
s38584.1	167.96	54.72	32.58
clma	282.60	75.04	26.55
Average			43.49

TABLE VI. COMPARISON OF THE CRITICAL PATH DELAY (CPD) BETWEEN THE GRADIENT BASED PLACEMENT ALGORITHM AND THE SIMULATED ANNEALING OF VPR

Netlist	VPR / ns	Gradient / ns	Relative / %
ex5p	77.43	136.99	176.92
tseng	54.05	86.71	160.42
apex4	115.96	145.58	125.54
misex3	80.33	94.23	117.31
alu4	81.17	115.10	141.79
diffeq	64.47	103.67	160.82
dsip	62.06	74.51	120.07
seq	120.73	117.41	97.25
apex2	96.96	121.85	125.67
s298	146.11	221.08	151.31
des	89.68	131.52	146.65
bigkey	62.56	84.72	135.42
frisc	133.22	278.46	209.02
spla	158.94	180.22	113.39
elliptic	138.25	275.43	199.22
ex1010	181.87	220.60	121.30
pdc	232.23	233.94	100.73
s38417	123.94	231.70	186.95
s38584.1	94.18	126.16	133.95
clma	231.10	300.13	129.87
Average			142.68

was outlined and will be addressed in future work.

As the current implementation of the gradient placer is executed only single-threaded, the next logic step would be to parallelize its execution to make it even faster. The calculation of the gradients could be executed in parallel on node level, and even large parts of the legalization (e.g., the assignment of nodes to the regions) could be parallelized. Hence, a multi-threaded implementation would be beneficial. Preliminary results show that even with a very simple multi-threaded implementation (executed on a multi-core processor), an acceleration of about 30 percent is possible. Furthermore, the authors are currently looking into a GPU-computing approach for the presented algorithm, which – at the point of writing – seems to be very promising.

Even though the gradient placement approach was shown to be comparably fast for large netlists, a more recent set of benchmarks like the one included in [14] – containing much larger netlists – could be used to underline the scalability of the approach.

REFERENCES

[1] T. Bostelmann, T. Thiemann, and S. Sawitzki, “Accelerating FPGA-placement with a gradient descent based algorithm,” in The Twelfth International Conference on Advances in Circuits, Electronics and Micro-electronics (CENICS), October 2019, pp. 13–18.

TABLE VII. COMPARISON OF THE TOTAL NET DELAY (TND) BETWEEN THE GRADIENT BASED PLACEMENT ALGORITHM AND THE SIMULATED ANNEALING OF VPR

Netlist	VPR / ns	Gradient / ns	Relative / %
ex5p	73.38	132.39	180.42
tseng	46.18	78.83	170.72
apex4	111.91	141.53	126.47
misex3	76.28	89.64	117.51
alu4	76.58	111.05	145.02
diffeq	57.14	96.35	168.62
dsip	59.64	72.10	120.88
seq	108.57	112.82	103.91
apex2	91.82	116.71	127.11
s298	137.69	212.11	154.05
des	87.27	128.57	147.32
bigkey	60.15	82.31	136.85
frisc	119.88	267.31	222.97
spla	153.80	175.08	113.84
elliptic	132.56	266.46	201.01
ex1010	177.27	215.46	121.55
pdc	226.54	228.80	100.99
s38417	117.70	226.01	192.02
s38584.1	91.77	122.66	133.66
clma	223.77	291.16	130.12
Average			145.75

[2] T. Kohonen, Self-Organizing Maps. Springer, 1995.

[3] T. Bostelmann and S. Sawitzki, “Improving FPGA placement with a self-organizing map,” in International Conference on Reconfigurable Computing and FPGAs (ReConFig), December 2013, pp. 1–6.

[4] T. Bostelmann and S. Sawitzki, “Improving the performance of a SOM-based FPGA-placement-algorithm using SIMD-hardware,” in The Ninth International Conference on Advances in Circuits, Electronics and Micro-electronics (CENICS), July 2016, pp. 13–15.

[5] T. Bostelmann, P. Kewisch, L. Bublies, and S. Sawitzki, “Improving FPGA-placement with a self-organizing map accelerated by GPU-computing,” International Journal On Advances in Systems and Measurements, vol. 10, no. 1 & 2, 2017, pp. 45–55.

[6] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” Science, vol. 220, May 1983, pp. 671–680.

[7] L. Ingber, “Adaptive simulated annealing (ASA): Lessons learned,” Control and Cybernetics, vol. 25, 1996, pp. 33–54.

[8] M. M. Atiqullah, “An efficient simple cooling schedule for simulated annealing,” in International Conference on Computational Science and Its Applications (ICCSA). Springer, 2004, pp. 396–404.

[9] E. Vansteenkiste, S. Lenders, and D. Stroobandt, “Liquid: Fast placement prototyping through steepest gradient descent movement,” in 2016 26th International Conference on Field Programmable Logic and Applications (FPL), August 2016, pp. 1–4.

[10] V. Betz and J. Rose, “VPR: A new packing, placement and routing tool for FPGA research,” in International Conference on Field Programmable Logic and Applications (FPL). Springer, 1997, pp. 213–222.

[11] S. Yang, “Logic synthesis and optimization benchmarks user guide version 3.0,” Microelectronics Center of North Carolina, Tech. Rep., 1991.

[12] M. Gort and J. H. Anderson, “Analytical placement for heterogeneous FPGAs,” in 22nd International Conference on Field Programmable Logic and Applications (FPL), August 2012, pp. 143–150.

[13] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, pp. 1–15.

[14] J. Luu et al., “VTR 7.0: Next generation architecture and CAD system for FPGAs,” ACM Trans. Reconfigurable Technol. Syst., vol. 7, no. 2, June 2014, pp. 6:1–6:30.

Energy Capture Methods by Piezoelectric Sensors and Applications

Irinela Chilibon
National Institute of Optoelectronics, INOE 2000
Bucharest-Magurele, Romania
e-mail: qilib@yahoo.com

***Abstract*—This paper presents different energy capture methods with application to the piezoelectric sensors, such as in small power electrical sources in applications. Piezoelectric material elements of sensors convert mechanical energy into electrical charges, due to the direct piezoelectric effect. Collecting energy from the environment is a major area of interest to the development of unconventional renewable energy sources and creates electrical energy, comparatively with the classic one. Potential energies sources from the environment could be successfully used. The photostrictive devices incorporate intelligent materials, illumination sensing and self-production of drive or control voltage.**

Keywords-piezoelectric sensor; supercapacitors; renewable energy; low-power, Perovskite Lead Lanthanum Zirconate Titanate, (PLZT).

I. INTRODUCTION

Capturing energy from the environment has shown considerable interest in recent research [1][2][3][4]. Piezoelectric material elements are able to convert mechanical energy into electrical charges, due to the direct piezoelectric effect. These electric charges could be used in low-power electronic circuitry that use electrically charged high-capacity capacitors. The mechanical sources could be vibrations, pulses and shocks. Piezoelectric sensors are suitable to convert the lower mechanical deformations directly into electricity, which can be used in small power electrical sources in applications as: supercapacitors, battery packs, interferometric lasers, etc.

Collecting energy from the environment is a major area of interest with extensive applications in the development of unconventional renewable energy sources. This study may clarify mechanisms for converting mechanical energy into electricity with a high efficiency of conversion. The *mechanical energy* could come from *environmental sources* such as: *wind, vibrations, shocks, rotary movements, wheel rotations, car engines, human breathing, blood flow, body movements, free or lost mechanical energy or acoustic and ultrasonic vibrations.*

This paper is organized as follows. Section II describes the piezoelectric materials and structures used. Section III discusses an overview of Energy Capture Methods in low-power electronics. Section IV presents a summary of low-power electronics.

The conclusions, future work and acknowledgement close the article.

II. PIEZOELECTRIC MATERIALS AND STRUCTURES

Piezoelectric materials, usually crystals or ceramics, have the capability to generate a small amount of current, when they are subjected to mechanical pressure, such as pushing, bending, twisting, and turning. Multiple such materials, placed near each other could increase the electrical energy.

The process of energy conversion in a piezoelectric material is based on the principle of the piezoelectric effect. The piezoelectric element stores the energy in two forms, as an electric field (electrical energy) and as a strain (mechanical energy). The *piezoelectric effect* exists in two domains, the first is the *direct piezoelectric effect* that describes the material's ability to transform mechanical strain into electrical charge, the second form is the converse effect, which is the ability to convert an applied electrical potential into mechanical strain energy, as shown in Figure 1. When a piezoelectric element is mechanically stressed, it generates a charge.

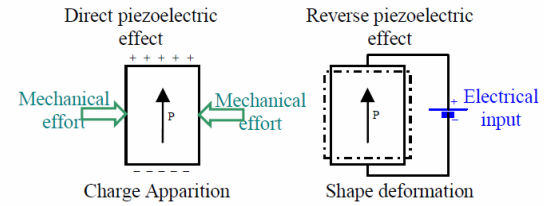


Figure 1. Electromechanical conversion via piezoelectricity phenomenon

The most common type of piezoelectric used in *power harvesting applications* is lead zirconate titanate, a piezoelectric ceramic, or piezoceramic, known as PZT. Although PZT is widely used as a power harvesting material, the piezoceramic's extremely brittle nature causes limitations in the strain that it can safely absorb without being damaged. Lee et al. [1] note that piezoceramics are susceptible to fatigue crack growth when subjected to high frequency cyclic loading. In order to eliminate the disadvantages of piezoceramic materials and improve upon their efficiency, researchers have developed and tested other, more flexible, piezoelectric materials that can be used in energy harvesting applications.

Another common piezoelectric material is *poly(vinylidene fluoride) (PVDF).* PVDF is a piezoelectric polymer that exhibits considerable flexibility when compared to PZT.

Mohammadi et al. [2] developed a *fiber-based piezoelectric (piezofiber) material* consisting of PZT fibers of various diameters (15, 45, 120, and 250 μm) that were aligned, laminated, and molded in an epoxy [3]. *Piezofiber power harvesting materials* have also been investigated by Churchill et al. [4], who tested a composite consisting of unidirectionally aligned PZT fibers of 250 μm diameter embedded in a resin matrix. It was found that when a 0.38 mm thick sample of 130 mm length and 13 mm width was subjected to a 180 Hz vibration that caused a strain of 300 με in the sample, the composite was able to harvest about 7.5 mW of power.

The last years have seen the birth of many new types of piezoelectric materials or transducers (PZT- lead zirconate titanate, PT-lead titanate, PVDF- polyvinylidene fluoride-trifluoroethylene, piezoceramic/polymer composites, Macro Fiber Composite - MFC, etc.). The schematic of the cross section of an Active Fiber Composite (AFC) actuator [5] is presented in Figure 2. If optimized geometrically, a piezoelectric generator associated with a well suited electronic is likely able to produce the 3 Watts standard required for the lighting system, with all the benefits that it provides.

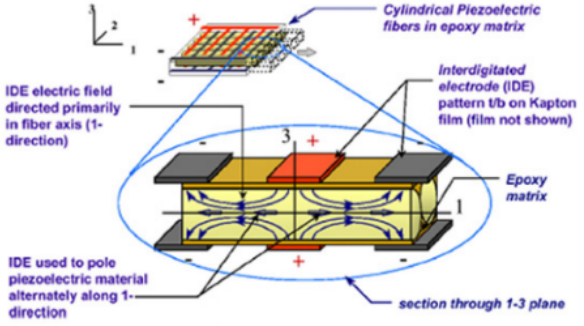


Figure 2. Schematic of the cross section of an Active Fiber Composite (AFC) actuator [5].

Piezoelectric materials exhibit the property that if they are mechanically strained, they generate an electric field proportional to the strain [6]. Conversely, when an electric field is applied, the material undergoes strain. These anisotropic relationships are described by the piezoelectric strain constant, *d*, which gives the relationship between applied stresses while the electro-mechanical coupling coefficient, *k*, describes the efficiency with which energy is converted between mechanical and electrical forms. This latter coefficient is important in determining the efficiency of a resonant generator since the overall efficiency of a piezo element clamped to a substrate and cyclically compressed at its resonant frequency is (1):

$$\eta = \frac{\frac{k^2}{2(1-k^2)}}{\frac{1}{Q}+\frac{k^2}{2(1-k^2)}} \quad (1)$$

where Q is the quality factor of the resonator. As Q becomes larger, the efficiency tends towards unity but, for typically achievable Q factors, the efficiency increases significantly for higher values of k.

Lee et al. [1][7] developed a PVDF film that was coated with poly(3,4-ethylenedioxy-thiophene)/poly(4-styrenesulfonate) [PEDOT/PSS] electrodes. They compared *the PEDOT/PSS coated films* to films coated with the inorganic electrode materials, *indium tin oxide (ITO)* and platinum (Pt). When subjected to vibrations of the same magnitude over varying frequencies, it was found that the films with Pt electrodes began to show fatigue crack damage of the electrode surface at a frequency of 33 kHz. The ITO electrodes became damaged when operating at a frequency of 213 Hz. The PEDOT/PSS film, however, ran for 10 h at 1 MHz without electrode damage. One can conclude that, by utilizing a more durable electrode layer, a piezoelectric device can operate under more strenuous conditions. This may give the device the ability to harvest more power throughout its lifespan; however, the exact effect of a stronger electrode layer may vary depending on the specific application.

III. OVERVIEW OF ENERGY CAPTURE METHODS IN LOW-POWER ELECTRONICS

Piezoelectric generators are appropriate to convert the smallest mechanical deformations directly into electrical energy. This solid-state effect is free of degradation in a wide operation range. Therefore, a very high lifetime and availability can be guaranteed.

Piezoelectric materials and transducers are available commercially. Thus, the new piezoelectric generators could be produced cost-efficiently in large quantities and easy exploitation. Nowadays, it is *possible of generating renewable electricity* using piezoelectric materials and transducers placed in special structures that allow amplification of the direct piezoelectric effect. The generation of electric charges is produced by mechanical motion action.

Several techniques have been proposed and developed to extract energy from the environment. The most common available sources of energy are: wind, solar, temperature and stress (pressure). In general, *vibration energy* could be converted into electrical energy using one of *three techniques: electrostatic charge*, *magnetic fields*, and *piezoelectric materials*.

A number of sources of harvestable ambient energy exist, including waste heat, vibration, electromagnetic waves, wind, flowing water, and solar energy. While each of these sources of energy can be effectively used to power remote sensors, the structural and biological communities

have placed an emphasis on scavenging vibrational energy with piezoelectric materials [8]. A piezoelectric material transforms electrical energy into mechanical strain energy, and likewise to transform mechanical strain energy into electrical charge [9].

As piezo energy harvesting has been investigated only since the late '90s, it remains an *emerging technology*. With the recent surge of microscale devices, piezoelectric power generation can provide a *convenient alternative to traditional power sources* used to operate certain types of *sensors/actuators*, *telemetry*, and Microelectromechanical systems *MEMS* devices. Scavenging energy from ambient vibrations, wind, heat or light could enable *smart sensors* to be functional indefinitely.

It is now necessary to develop the structures of materials with high piezoelectric coefficients, an optimal architecture to increase the electrical efficiency of specialized devices for alternative energy production, and cheap to replace traditional energy sources part.

Now it is necessary to develop the structures of materials with high piezoelectric coefficients, an optimal architecture for increasing the electrical efficiency of specialized devices for the production of cheap alternative energy to replace the part of traditional energy sources.

Advances in low-power electronics and in energy harvesting technologies have enabled the conception of truly self-powered devices [10]. Cantilevered piezoelectric energy harvesters have been investigated in the literature of energy harvesting [6][16].

The concept of *"harvesting"* is recent and involves capturing the energy normally lost around a system and converting it into electricity that can be used to extend the life of the system or to provide an endless source of energy to a system [11].

IV. OVERVIEW OF LOW-POWER ELECTRONICS

The methods of accumulating and storing the energy generated, until sufficient power has been captured, is the *key to develop completely self-powered systems***.** Piezoelectric transduction has received great attention for vibration-to-electric energy conversion over the last five years [12]. Future applications may include high power output devices (or arrays of such devices) deployed at remote locations to serve as reliable power stations for large systems, *wearable electronics***.**

Among those *challenges* is the electronic circuitry needed to capture, accumulate and store energy from energy harvesting energy sources. The circuitry must then switch the power from an energy storage device and then supply it to the application. In general, electricity can be stored in various electronic components, such as: capacitors, supercapacitors or batteries. Piezoelectric generators are appropriate to convert the smallest mechanical deformations directly into electrical energy. Among *alternative energy sources* remember: water energy, solar, wind, heat, etc. It needs to find other methods and possible ways of using energy sources such as mechanical energy converted into electrical energy.

Many conventional systems consist of a single piezoceramic in bending mode (*unimorph*) or two bonded piezoelectric in bending mode (*bimorph*), but upon the experimental validating model had 4.61 % maximum error [9]. Some structures can be tuned to have two natural frequencies relatively close to each other, resulting in the possibility of a *broader band energy harvesting system* [12] [13]. The energy produced by these materials is in many cases far too small to directly power an electrical device [11][14]. Recent studies present the ability to take the energy generated through the vibration of a piezoelectric material was shown to be capable of recharging a discharged battery [9].

Bimorph actuators consist of two independent flat piezoelectric elements, stacked on top of the other. By driving one element to expand while contracting the other one, the actuator is forced to bend, creating an out-of plane motion and vibrations [15]. *Cantilevered* piezoelectric energy harvesters have been investigated in the literature for energy harvesting [12]. A more attractive configuration is to form the piezoceramic into a cantilever arrangement, as shown in Figure 3 where layers of piezoceramics are bonded to a substrate, typically made from a suitable metal. This structure allows a lower resonant frequency to be achieved while producing large strains in the piezoceramic. Where two layers of piezo material are used, the structure is referred to as a bimorph. In this case, the piezo layers may either be connected in series or parallel. If only a single piezo layer is used, the structure is referred to as a *unimorph* [6].

A. Piezoelectric devices

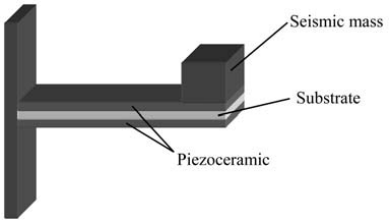


Figure 3. Piezoceramic cantilever resonator [6].

Figure 4 presents a power generator array prototype, realized by small cantilevers of different lengths, in order to obtain a larger broadband [16].

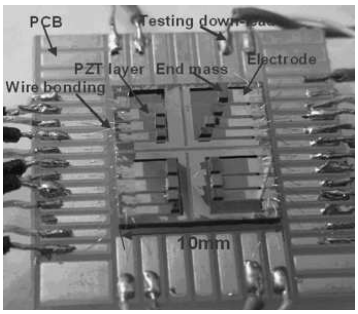


Figure 4. Picture of power generator array prototype [16].

The main steps of fabrication process of micro piezoelectric power generator is presented in the Figure 5, namely: (1) Functional films preparation: SiO_2/Ti/Pt/PZT /Ti/Pt, (2) functional films pattern, (3) silicon slot etching by RIE, (4) back silicon deep etching by KOH solution, (5) cantilever release by RIE, and (6) metal mass micro fabrication and assemblage [17], as shown in Figure 5.

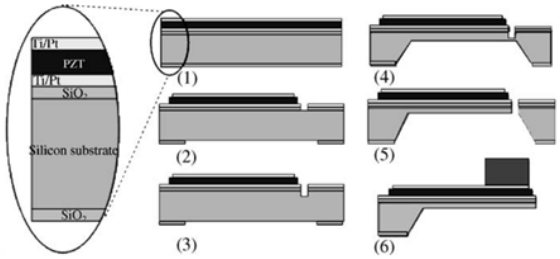


Figure 5. Fabrication process of micro piezoelectric power generator [17].

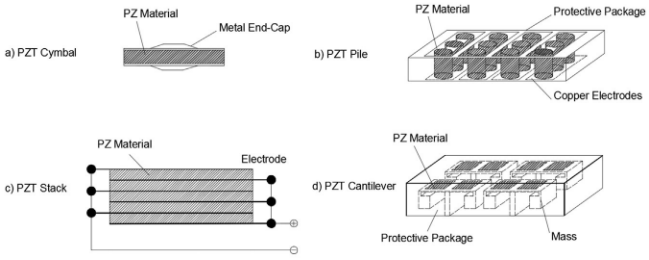


Figure 6. Types of PZT energy harvesters in pavement [18].

Studies about the piezoelectric effects for energy harvesting pavement were made by [18], using different structures of Piezoelectric Sensors, like: PZT cymbal, PZT pile, PZT stack and PZT cantilever, as shown in Figure 6.

B. Energy harvesting piezoelectric circuitry

Piezoelectric generators are appropriate to convert the smallest mechanical deformations directly into electrical energy. This solid-state effect is free of degradation in a wide operation range. A vibrating piezoelectric device differs from a typical electrical power source in that it has capacitive rather than inductive source impedance, and may be driven by mechanical vibrations of varying amplitude, as shown in Figure 7. A PZT disc for example, compressed between two metal surfaces will never be able to expand in the radial direction as readily as would a long, thin cylinder, which is only constrained at its ends and assumes a barrel shape on radial expansion. So, the way in which the material is mounted will directly affect the energy conversion per unit volume. The general rule therefore is to allow the PZT body some freedom to expand radially since charge generation is directly coupled to deformation.

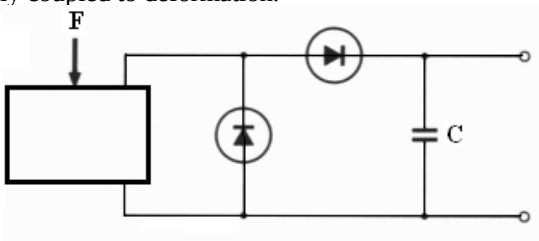


Figure 7. PZT element generator with 2 diodes as DC converter and a parallel capacitor C.

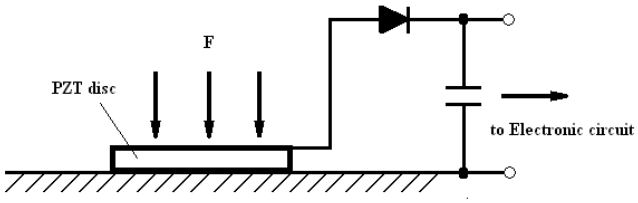


Figure 8. Electronic circuit with PZT disc strained by F force, and one diode DC converter.

The principle of charge generation by a PZT disc to an electronic circuit performance are the shape of the PZT transducer, the manner in which the transducer is mounted and, of course, the nature of the electrical load, as shown in Figure 8.

Typical energy harvesting circuitry consists of voltage rectifier, converter and storage, as shown in Figure 9.

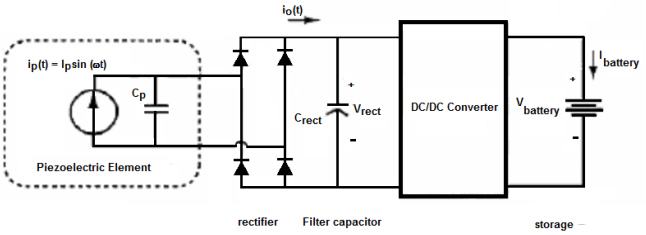


Figure 9. Typical energy harvesting circuitry

Energy harvesting sources that generate power from ambient sources present problems in generating a predictable flow of electricity for the operation of electronic circuits. At times these sources generate zero power. At other times they generate trace amounts of power that are unusable. Then there are times when the power generated is so great that a charge from an energy harvesting source could burn out the circuitry. Therefore, it should be used electronics with energy harvesting intelligent piezoelectric transducer. In Wireless Sensor Networks (WSNs), one of major hurdles is the limited battery power that is unable to meet long-term energy requirement. Energy harvesting, conversion of

ambient energy into electrical energy, has emerged as an effective alternative to powering WSNs [19].

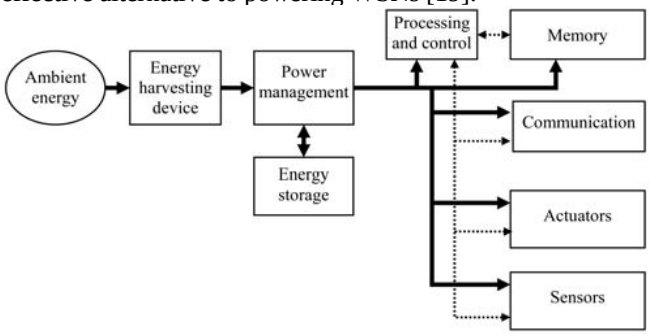


Figure 10. A generic sensor network node with energy harvesting device [6].

The idea of harvesting ambient energy from the immediate surroundings of the deployed sensors is to recharge the batteries and to directly power the sensor nodes. The power consumed by a network node can be split between the various functions it has to perform, as shown in Figure 10.

A low power piezoelectric generator, having a PZT element was realized in order to supply small electronic elements, such as optoelectronic small devices, LEDs, electronic watches, small sensors, interferometry with lasers or Micro-electro-mechanical System (MEMS) array with multi-cantilevers **[20].** A Set-up for the experimental work was realized **[20]**, Figure 11.

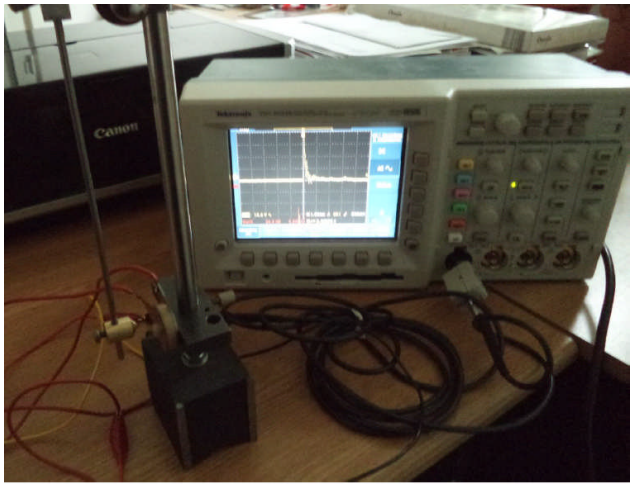


Figure 11. Set-up for the experimental work **[20]**.

A PZT element hit by mechanical shocks can generate high voltage electrical sparks that processed by electronic circuits is stored in special batteries.

Energy harvesting technologies have been explored by researchers for more than two decades as an alternative to conventional power sources (e.g. batteries) for small-sized and low-power electronic devices [21].

The vibrations energy harvesting principle using piezoelectric materials is illustrated in Figure 12. The conversion chain starts with a mechanical energy source, like vibrations which are converted into electricity via piezoelectric element. The electricity produced is thereafter formatted by a static converter before supplying a storage system or the load (electrical device).

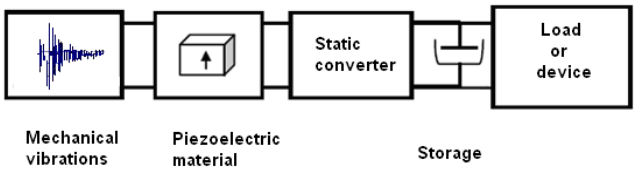


Figure 12. General diagram of a generator, based vibrations energy harvesting using piezoelectric material

Advances in low-power electronics and in energy harvesting technologies have enabled the conception of truly **self-powered devices** [10].

Band gap and Young's modulus of elasticity of one-dimensional nanostructures for ZnO were studied. Also, how these properties make it a superior material for piezoelectric energy harvesting [22].

Also, the principle that this material is superior to other piezoelectric energy harvesting has been established. Other materials have been explored as being good for building such energy harvesters. A single nanowire can function as a nano-transducer and generate piezoelectricity, and integrating multiple of these nanowires into arrays on a single substrate could improve the output power. The use of polymers and ceramics has been studied for the same energy harvester. A nutshell, for piezoelectric energy harvesting and nanostructures could be manufactured and assembled to work as a self-powered nano / micro-structured device [23].

Piezoelectric materials are capable of converting irregular and low-frequency mechanical vibration into electricity and useful for nanogenerators. Energy harvesting devices can convert energy in other forms into the electricity but have limited energy storage capabilities.

Polarized Polyvinylidene Fluoride (PVDF) films were successful used in supercapacitors manufacture. A piezo-supercapacitor was made from a sandwich of a PVDF layer enclosed between two functionalized carbon cloths (FCC) electrodes. As result, a voltage potential was generated across the piezoelectric PVDF under external mechanical vibration [24].

C. The photostriction phenomenon

In certain ferroelectrics, a constant electromotive force is generated with exposure to light, and a photostrictive strain results from the coupling of this bulk photovoltaic effect to converse piezoelectricity. A photostrictive actuator is a fine example of an intelligent material, incorporating illumination sensing and self-production of drive/control voltage together with final actuation [25]. When a violet light is irradiated to one side of the Perovskite Lead Lanthanum Zirconate Titanate (PLZT) bimorph, enormous photovoltage of 1 kV mm^{-1} is generated, causing a bending motion. The displacement of the tip of a bimorph device

having a length of 20 mm, and a thickness of 0.4 mm was 150 μm, within a response time of 1 s.

V. CONCLUSION AND FUTURE WORK

Collecting energy from the environment is a major area of interest with extensive applications in the development of unconventional renewable energy sources. This study clarifies mechanisms for converting mechanical energy into electricity at high efficiency conversion.

Energy harvesting is an attractive concept because so many energy sources such as light, heat, and mechanical vibration that exist in our ambient living could be converted into usable electricity. The technical progress in the *field of extremely low energy electronics* opens up the chance to use *harvested energy* from the environment. Most piezoelectric electricity sources produce power on the order of milliwatts, too small for system application, but enough for hand-held devices such as some commercially available self-winding wristwatches. The methods of accumulating and storing the energy generated, until sufficient power has been captured, is the *key to develop completely self-powered systems***.**

WSNs are crucial in supporting continuous environmental monitoring, where sensor nodes are deployed and must remain operational to collect and transfer data from the environment to a base-station. Further development of such energy circuitry and piezoelectric materials with various structures will facilitate the progression of power harvesting methods from a research topic to a useable technology in practical devices.

The photostriction phenomenon appears in certain ferroelectrics, when a constant electromotive force is generated with exposure to light, and a photostrictive strain results from the coupling of this bulk photovoltaic effect to converse piezoelectricity.

The PVDF material has good properties of piezoelectricity and mechanical flexibility, and demonstrated as a wearable and flexible energy generator. Polarized PVDF films were successful used in supercapacitors manufacture.

ACKNOWLEDGMENT

This work was financed by the Romanian Ministry of Research and Innovation (MCI), 2019 Core Program, Contract nr. PN19-18.01.02/2019, Stage I, Stage II, and the 19 PFE-PDI/2018 Project, Stage II/2019.

REFERENCES

[1] C. S. Lee, J. Joo, S. Han, J. H. Lee, and S. K. Koh, "Poly(vinylidenefluoride) transducers with highly conducting poly(3,4-ethylenedioxythiophene) electrodes," Proc. Int. Conf. on Science and Technology of Synthetic Metals, vol. 152, pp. 49–52, 2005.

[2] F. Mohammadi, A. Khan, and R. B. Cass, "Power generation from piezoelectric lead zirconate titanate fiber composites," Proc. Materials Research Symp., pp. 736, 2003.

[3] A. A. Bent, N. W. Hagood and J. P. Rodgers, "Anisotropic actuation with piezoelectric fiber composites," J. Intell. Mater. Syst. Struct., vol. 6, pp. 338–349, 1995.

[4] D. L. Churchill, M. J. Hamel, C. P. Townsend, and S. W. Arms, Strain energy harvesting for wireless sensor networks," Proc. Smart Struct. and Mater. Conf., Proc. SPIE 5055, pp. 319, 2003.

[5] W. K. Wilkie et al., "Low-cost piezocomposite actuator for structural control applications," Proc.7th Int. Symp., 2000.

[6] J. M. Gilbert and F. Balouchi, "Comparison of Energy Harvesting Systems for Wireless Sensor Networks," International Journal of Automation and Computing, vol. 05(4), pp. 334-347, Oct. 2008.

[7] C. S. Lee, J. Joo, S. Han, and S. K. Koh, "Multifunctional transducer using poly(vinylidene fluoride) active layer and highly conducting poly(3,4-ethylenedioxythiophene) electrode: actuator and generator," Appl. Phys. Lett., vol 85, pp. 1841–3, 2004.

[8] S. R. Anton and H. A. Sodano, "A review of power harvesting using piezoelectric materials (2003–2006)", Smart Mater. Struct., vol. 16, pp. R1–R21, 2007, doi:10.1088/0964-1726/16/3/R01.

[9] H. A. Sodano, D. J. Inman, and G. Park, "A Review of Power Harvesting from Vibration using Piezoelectric Materials," The Shock and Vibration Digest, pp. 197-205, May 2004.

[10] M. Lallart, S. Priya, S. Bressers, and D. J. Inman, "Small-scale piezoelectric energy harvesting devices using low-energy-density sources," Journal of the Korean Physical Society, vol. 57(41), pp. 947-951, Oct 15 2010.

[11] H. Sodano and D. Inman, "Generation and Storage of Electricity from Power Harvesting Devices," Journal of Intelligent Material Systems and Structures, vol. 16(1), pp. 67-75, Jan. 2005, doi: 10.1177/1045389X05047210.

[12] A. Erturk and D. J. Inman, "An experimentally validated bimorph cantilever model for piezoelectric energy harvesting from base excitations," Smart Mater. Struct., vol. 18, pp. 025009 (18pp), 2009, doi:10.1088/0964-1726/18/2/025009.

[13] A. Erturk and D. J. Inman, "A Distributed Parameter Electromechanical Model for Cantilevered Piezoelectric Energy Harvesters", J. Vib. Acoust, vol. 130(4), pp. 041002 (15 pages), June 2008, doi:10.1115/1.2890402.

[14] N. M. White, P. Glynne-Jones, and S. P. Beeby, "A novel thick-film piezoelectric micro-generator," Smart Materials & Structures, *10*(4), pp. 850-852, 2001.

[15] I. Chilibon, C. Dias, P. Inacio, and J. Marat-Mendes, "PZT and PVDF bimorph actuators," Journal of Optoelectronics and Advanced Materials, vol. 9**,** Issue 6, pp. 1939-1943, Jun 2007, ISSN: 1454-4164.

[16] J-Q. Liu et al., "A MEMS-based piezoelectric power generator array for vibration energy harvesting," Microelectronics Journal, vol. 39, pp. 802–806, 2008.

[17] H-B Fang et al., "Fabrication and performance of MEMS-based piezoelectric power generator for vibration energy harvesting," Microelectronics Journal, vol. 37, pp. 1280–1284, 2006.

[18] L. Guo and Q. Lu, "Potentials of piezoelectric and thermoelectric technologies for harvesting energy from pavements," Renewable and Sustainable Energy Reviews, vol. 72, pp. 761-773, May 2017, doi: 10.1016/j.rser.2017.01.090.

[19] K. Singh and S. Moh, "Comparative Survey of Energy Harvesting Techniques for Wireless Sensor Networks," Advanced Science and Technology Letters, vol.142, pp. 28-33, GDC 2016, http://dx.doi.org/10.14257/astl.2016.142.05.

[20] I. Chilibon, Piezoelectric devices for generating low power, *Proc. SPIE* 10010, Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies VIII, Vol. SPIE 10010, Article Number: UNSP 100101H, (Dec. 14, 2016), doi: 10.1117/12.2248729

[21] M. Safaei, H. A. Sodano and S. R. Anton, "A review of energy harvesting using piezoelectric materials: state-of-the-art a decade later (2008–2018)", Smart Materials and Structures,

Smart Mater. Struct., vol. 28, 113001, (62 pp), 2019, https://doi.org/10.1088/1361-665X/ab36e4

[22] R. Agrawal, B. Peng, E. E. Gdoutos, and H. D. Espinosa "Elasticity size effects in ZnO nanowires – a combined experimental-computational approach," Nano Lett. 8(11): 3668–3674, 2008.

[23] I. Dakua, and N. Afzulpurkar, "Piezoelectric Energy Generation and Harvesting at the Nano-Scale: Materials and Devices, Review Article," Nanomater. Nanotechnol., vol. 3, Art. 21:20, August 2013. https://journals.sagepub.com/doi/pdf/10.5772/56941

[24] R. Song, et al., "A Rectification-Free Piezo-Supercapacitor with a Polyvinylidene Fluoride Separator and Functionalized Carbon Cloth Electrodes," J. Mater. Chem. A, July 2015, doi: 10.1039/C5TA03349G

[25] K. Uchino, "Focus Issue Review: Glory of piezoelectric perovskites," Sci. Technol. Adv. Mater. 16 046001 (16 pages) , 2015, doi:10.1088/1468-6996/16/4/046001