ANFIS-based Wireless Sensor Network (WSN) Applications for Air Conditioner Control

Yi-Jen Mon[1], Chih-Min Lin[2], Imre J. Rudas[3]

[1] Department of Computer Science and Information Engineering, Taoyuan Innovation Institute of Technology, Chung-Li, Taoyuan, 320, Taiwan
e-mail: monbuy@tiit.edu.tw

[2] Department of Electrical Engineering, Yuan Ze University, Chung-Li, Taoyuan, 320, Taiwan, e-mail: cml@saturn.yzu.edu.tw

[3] Óbuda University, Bécsi út 96/B, H-1034 Budapest, Hungary
e-mail: rudas@uni-obuda.hu

Abstract: The adaptive network fuzzy inference system (ANFIS)-based wireless sensor network (WSN) is designed to serve as a monitor of controller of an indoor air-conditioning system. The WSN comprises sensors to monitor the temperature of the indoor space and the ANFIS controller is used to control the fans in order to obtain energy-saving benefits. By using the application programming interface (API) of WSNs, many applications have been developed. The experimental results demonstrate that good data transference and control performance have been achieved.

Keywords: wireless sensor network; ANFIS; air conditioner control

1 Introduction

In this paper, an adaptive network fuzzy inference system (ANFIS)-based wireless sensor network (WSN) control system is developed. A reduction in energy consumption is the main challenge of the WSN [1]. The IEEE 802.15.4/ZigBee standard has been proposed by the ZigBee Alliance to develop standards for cost-effective and low-power consumption WSNs [2-10]. By using the lower cost, lower power consumption ZigBee WSN and efficient ANFIS controller, many control applications can be achieved through the network. Recently, many applications of indoor, outdoor and mobile devices, such as home automation and security, vehicle space automation, consumer products, health-care, environmental monitoring and indoor location identification, etc., have been developed to improve human quality of life. In particular, these WSNs can provide ideal networking solutions for lower cost and simple installations [1, 2].

The ZigBee, which builds upon the IEEE 802.15.4 standard, defines the physical layers and media access control (MAC) layers for lower cost, lower rate, personal area networks (PAN). There are three types of ZigBee network topologies defined. These networks topologies are star, tree and mesh network topologies, which are provided by a framework of application programming interfaces (API) in the application layer [2-10].

The ANFIS [11] was proposed many years ago and is widely used in research works. The ANFIS reveals an efficient learning network and its applications can be found in many works in the literature [12-14]. In this paper, the ANFIS controller is used to control the motors of fans for indoor temperature control. The ANFIS controller is a user-friendly algorithm and serves as a method for inducing many fuzzy if-then rules with suitable membership functions to generate the relationship fuzzy associated memory (FAM) pairs of inputs/outputs (I/O) and reasonable fuzzy rules so as to achieve a fuzzy inference system (FIS). A reliable controller can be designed based on this FIS. *MATLAB*™ Ltd. has provided very useful and user-friendly tools for engineers to design this ANFIS controller [15].

In the solution reported in this paper, by using the data measured via the WSN, data transference and the fan's motor control are performed by the proposed ANFIS-based WSN control methodology. This method has great benefits, such as reducing the power consumption of the indoor space, improving the efficiency of air-conditioning and saving energy.

2 Preliminary of ZigBee WSN

Regarding the MAC layer of ZigBee and integrated peripherals of the WSN microcontroller (WMCU), the Application Queue APIs (AQA) of the WSN microcontroller are provided a queue-based interface between applications and both the ZigBee stack and the hardware drivers. These AQAs are used to deal with many interrupts coming from the MAC layers. There are three types of interrupt implemented by using APIs. The first type is used for the MAC data services, the second type is used for the MAC management services and the third type is used for handling interrupts of hardware drivers [16, 17].

The basic type of network topology is the star topology, which comprises a central personal area network (PAN), which is called 'coordinator', surrounded by the other nodes of the network, which are called 'end device'. The tree network topology has an implicit structure based on parent-child relationships. In the mesh network topology, all devices can communicate directly and can be identical in an ad-hoc network [16, 17]. In this paper, the star topology of a WSN is used to develop the air conditioner controller. Data transference between network nodes can be searched or found by any request. The data transference methods are shown in

Fig. 1. When transferring data from a coordinator to a node, the node might not always be ready to receive the data. In this case, the node will be requested to receive data frame. Afterwards, the 'Acknowledgment' message is sent from coordinator. Finally, the end device will send the 'Acknowledgment' message to the coordinator once the data transferring has finished. [16, 17].

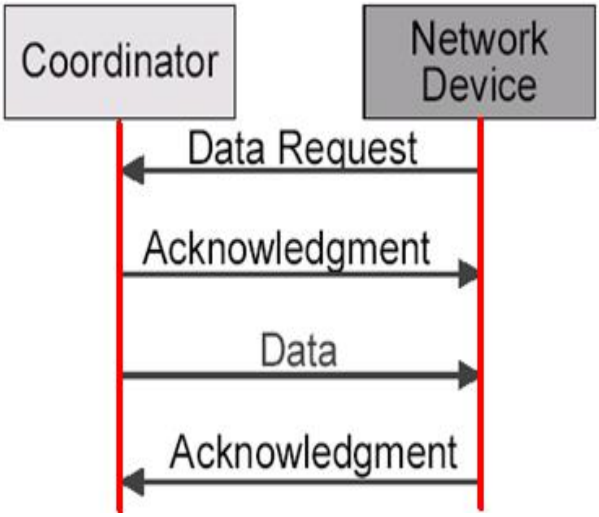


Figure 1
The diagram of data transference method of a WSN

3 ANFIS-based WSN Controller Design

The algorithm of ANFIS is based on a least-squares estimation (LSE) and back propagation gradient descent methods to identify the membership functions' parameters so as to achieve fuzzy inference systems (FIS). A neural network structure of ANFIS is shown in Fig. 2. This methodology requires a pair of input/output data to train the FIS membership function parameters. The developed method of ANFIS is briefly described as follows [11]:

$$R_i : If\ x_1\ is\ A_{i1}\ \ ...and\ x_n\ is\ A_{in}$$

$$then\ u_i = p_{i1}x_1 + .. + p_{in}x_n + r_i \tag{1}$$

where R_i denotes the ith fuzzy rules, i=1, 2,..,r; A_{ik} is the fuzzy set in the antecedent part associated with the kth input variable at the ith fuzzy rule, and $p_{i1},...,p_{in}$, r_i are the fuzzy consequent parameters.

The defuzzification of the output u can be calculated by the method of averaged weight as follows:

$$u = \frac{w_1}{w_1 + .. + w_n} u_1 + .. + \frac{w_n}{w_1 + .. + w_n} u_n$$

$$= \overline{w}_1 u_1 + .. + \overline{w}_2 u_n \qquad (2)$$

where w_i is the ith node output firing strength of the ith rule, and $\overline{w}_1 = \frac{w_1}{w_1 + .. + w_n}, .., \overline{w}_n = \frac{w_n}{w_1 + .. + w_n}$.

Based on the reason of Takagi-Sugeno (T-S) type fuzzy inference system, the output of ANFIS can be calculated as $u_i = p_{i1}x_1 + .. + p_{in}x_n + r_i$, then Eq. (2) can be expressed as

$$\begin{aligned} u &= \overline{w}_1 u_1 + .. + \overline{w}_n u_n \\ &= (\overline{w}_1 x_1) p_{i1} + .. + (\overline{w}_1 x_n) p_{in} + (\overline{w}_1) r_1 \\ &+ \\ &\vdots \\ &+ (\overline{w}_n x_1) p_{i1} + .. + (\overline{w}_n x_n) p_{in} + (\overline{w}_n) r_n . \qquad (3) \end{aligned}$$

The ANFIS's algorithm can be applied directly to Eq. (3) [11]. In the forward direction of the ANFIS algorithm, input signals go forward until layer 4 of Fig. 2 and the consequent parameters p_{i1}, p_{i2}, r_i are changed. In the backward direction, the error values are feedbacked and used to update the premise parameters x_1, x_2. When all the values of the parameters are changed, the membership functions are also modified; thus every membership functions of A_{i1} and A_{i2} are induced.

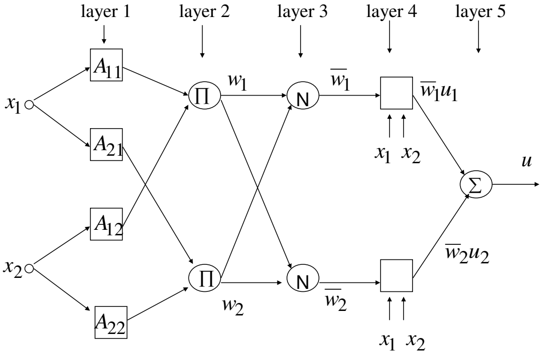


Figure 2
A two-input one-output ANFIS architecture

4 Experimental Results

The program of the WSN is developed on free software called Code::Blocks. First, the program of the coordinator is developed and then the program of the end device is developed. Every network must have one and only one PAN coordinator, and one of the tasks in setting up a network is to select and initialise this coordinator. The network setup process is shown in Fig. 3. The main program of the coordinator and the end device are developed in C language. The architecture of the coordinator software is shown in Fig. 4, especially the API of 'vProcessEventQueues', which is the most important subroutine of this system. For the configuration program, the personal area network identification (PAN-ID) of every end device must be set adequately. The development board is produced by *Fontal Technology Inc., Taiwan*. This is a high-power ZigBee Kit (named FT-6200). It can provide all the software tools and hardware required to obtain first-hand experience with the WSN. The entry-level kits contain one base development board (BDB) and one sensor development board (SDB). Each board is equipped with a high-power IEEE 802.15.4/ZigBee RF module based on the JN-5121 WMCU, which provides a much higher coverage range with a 2.4 GHz RF antenna, which has the IPEX connector for easier mechanical design than the normal power RF module. For the I/O expansion ports, it has 10 useful pins of general purpose input/output (GPIO), which include the universal asynchronous receiver/transmitter (UART), analogue-to-digital converter (ADC), digital-to-analogue converter (DAC) and Comparator. The sensor development board features temperature and humidity sensors [16, 17]. The development board is shown in Fig. 5.

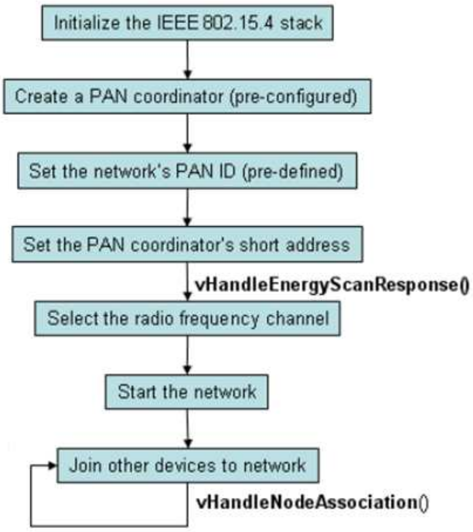

Initialize the IEEE 802.15.4 stack
Create a PAN coordinator (pre-configured)
Set the network's PAN ID (pre-defined)
Set the PAN coordinator's short address
vHandleEnergyScanResponse()
Select the radio frequency channel
Start the network
Join other devices to network
vHandleNodeAssociation()

Figure 3

Diagram of network setup process

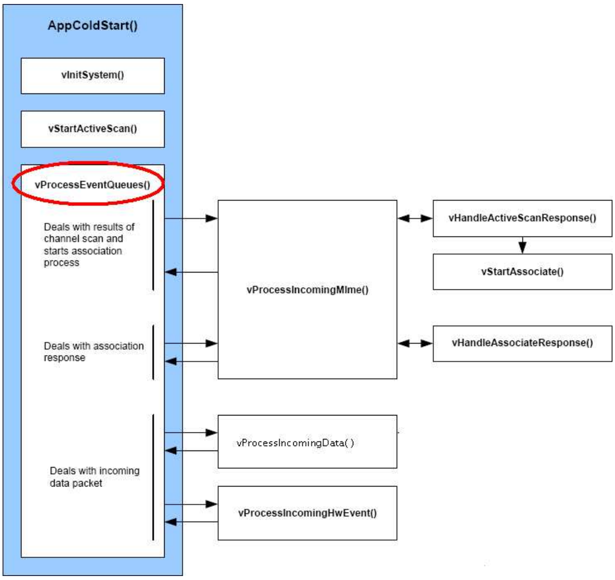

AppColdStart()
vInitSystem()
vStartActiveScan()
vProcessEventQueues()
Deals with results of channel scan and starts association process
Deals with association response
Deals with incoming data packet
vProcessIncomingMlme()
vHandleActiveScanResponse()
vStartAssociate()
vHandleAssociateResponse()
vProcessIncomingData( )
vProcessIncomingHwEvent()

Figure 4
Diagram of coordinator software architecture of WSN

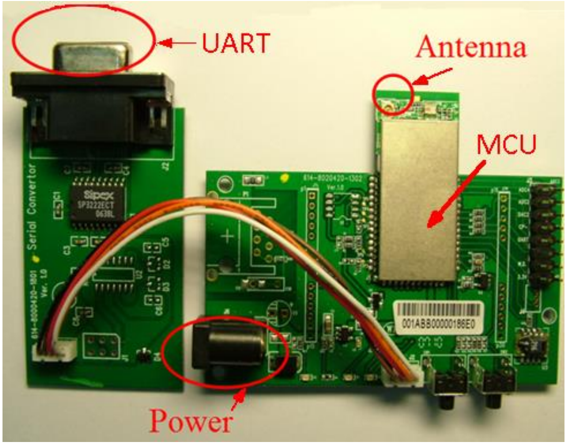


Figure 5
Development boards of WSN

For the software, *Jennic Technology Inc.* also provides free Application Programming Interface (API) packages to the peripheral devices on the JN5121 and JN513x single-chip IEEE 802.15.4/ZigBee compliant wireless microcontrollers. This is known as the Integrated Peripherals API. It details the calls that might be made through the API in order to set up, control and respond to events generated by the peripheral blocks, such as the UART, GPIO lines and Timers, among others.

The software invoked by this API is present in the on-chip ROM. This API does not include support for the ZigBee WSN MAC hardware built into the device; this hardware is controlled using the MAC software stack that is built into the on-chip ROM [16, 17].

ZigBee can be used with different sensors, such as: in-vehicle or home automation, security management, industrial, environmental controls and personal medical care. In this paper, the ZigBee WSNs are used to design the indoor air conditioner controller by means of the ANFIS-based WSN control methodology. The design concept diagram of the ANFIS-based WSN control is shown in Fig. 6. A star topology network is used in this paper. By using UART, the data can be displayed in the LCD of different end devices sensors located in four corners of the indoor space. Monitoring of the temperature is one of the main experimental aims; the temperature sensors on the end devices transmit data to the coordinator, which are then also displayed in the LCD through UART. An actual implementation of the air conditioner control in the laboratory is shown in Fig. 7. By simulation, the ANFIS-based controller described in section 3 is developed. Four end devices are used to measure temperature and then transmit that data to the coordinator. After manipulation by the ANFIS-based WSN controller, the fuzzy values are returned to the end devices located in the four corners of the indoor spaces to control the fans and adjust the temperature. The results of the ANFIS design are shown in Fig. 8. Especially from Fig. 8(b) of this simulation, reasonable fan speeds normalized from 0 to 1 are achieved. In this condition, for example, WSN1 has the highest temperature; thus, fan1 will be speeded up the most because it has the highest fuzzy number of 0.941. In this experiment, the GPIO of every end devices is used to perform the digital-to-analogue (D/A) transformation to send the signal to drive the motor of the fan. The results of the ten-hour continuous simulation are presented in Fig. 9. It can be clearly verified that fan 1 will be fully active from the 4th hour to the 5th hour, as shown in Fig. 9(b). From these empirical tests and simulations, the WSN control of temperature monitoring is successfully established and good performance of the fan motor control is achieved.

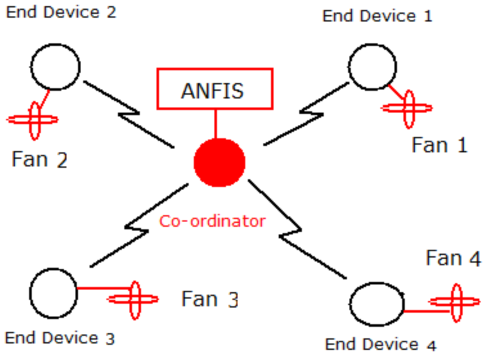


Figure 6
Design concept diagram of ANFIS-based WSN for air conditioner control

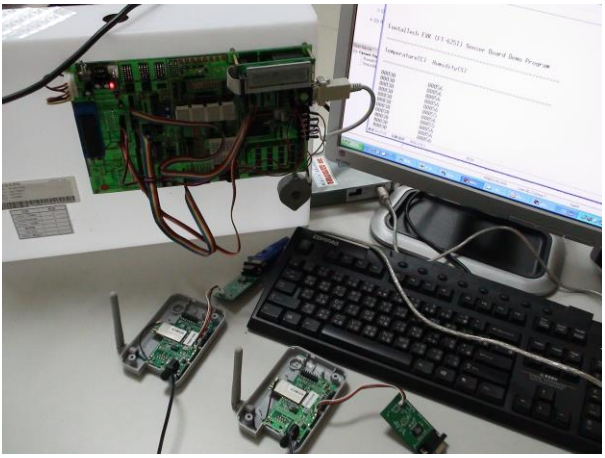
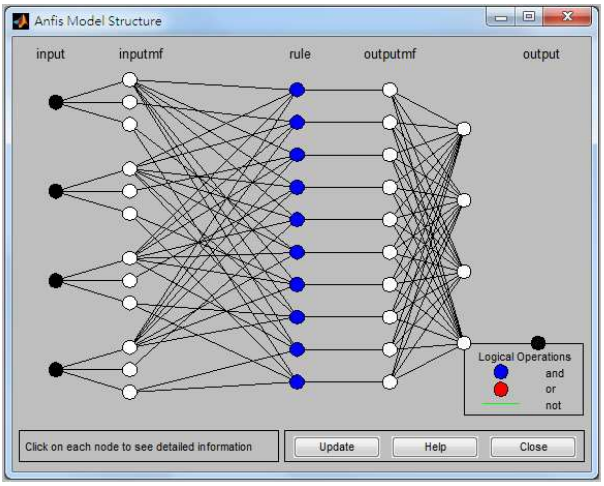
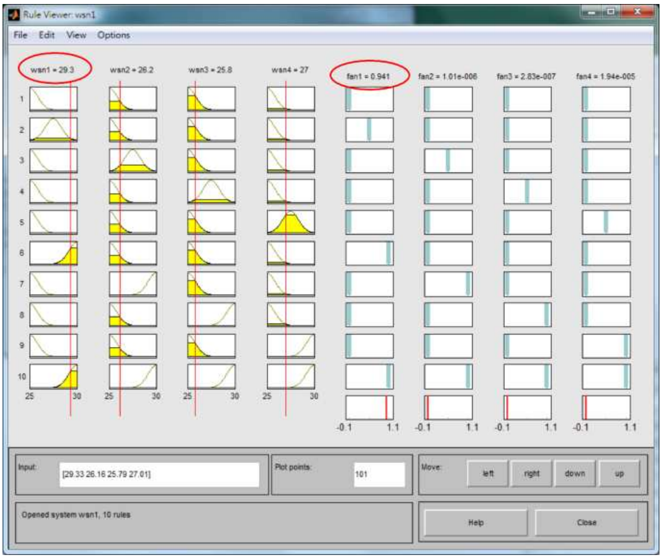


Figure 7
Implementation diagram of WSN for indoor temperature monitor



Click on each node to see detailed information

Figure 8 (a)
ANFIS structure diagram



Rule Viewer: wsn1

wsn1 = 29.3 wsn2 = 26.2 wsn3 = 25.8 wsn4 = 27 fan1 = 0.941 fan2 = 1.01e-006 fan3 = 2.83e-007 fan4 = 1.94e-005

Input: [29.33 26.16 25.79 27.01]

Plot points: 101

Opened system wsn1, 10 rules

Figure 8 (b)
ANFIS inference result diagram

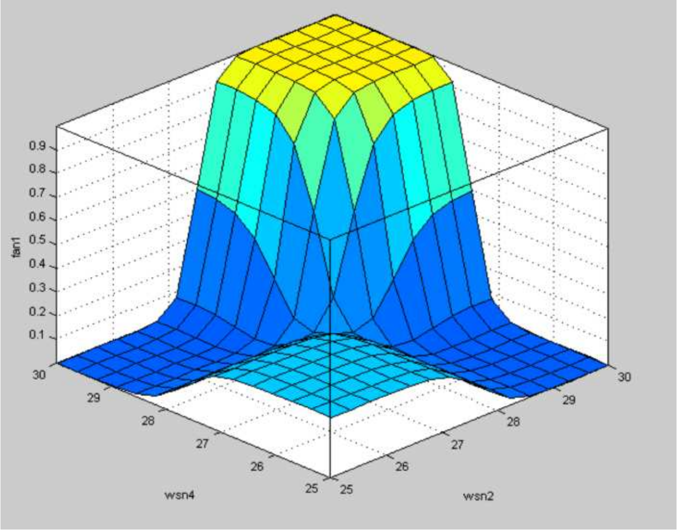


Figure 8 (c)
One of the ANFIS inference surface diagrams

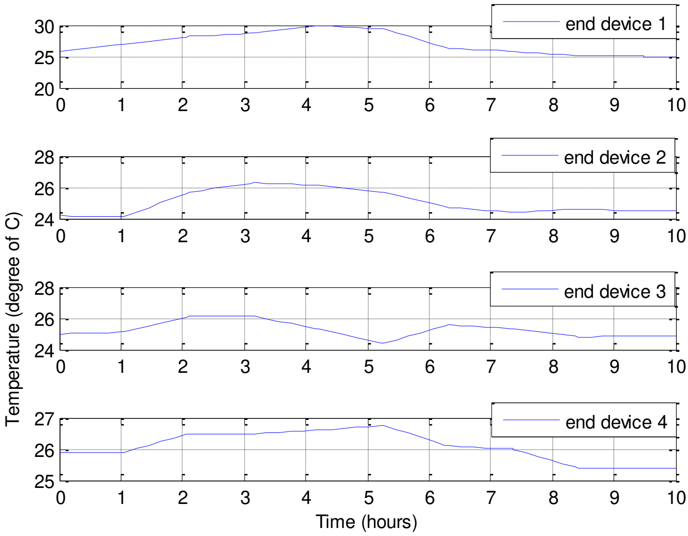


Figure 9 (a)

Diagram of measured temperatures of four corners indoor

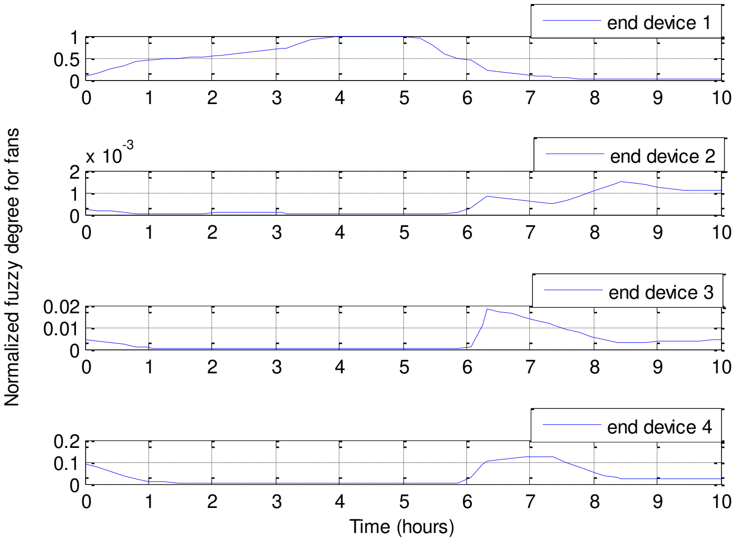


Figure 9 (b)

Diagram of simulation results of four corners indoor

Conclusions

A design method for the control of indoor an air conditioner by using the ANFIS-based WSN is proposed. This paper has successfully demonstrated the application of the WSN to monitor the indoor temperature. Physical verifications and simulations are also successfully demonstrated to show that satisfactory performance of the ANFIS-based WSN control of the motors of the fans, of the data collection and of temperature monitoring.

Acknowledgments

This paper is partially funded by teacher's research project of Taoyuan Innovation Institute of Technology.

References

[1] FT-625x Development Kits User Guide (Tradition Chinese version) 2012 (http://surewin.com.tw)

[2] R. Belbachir, Z. M. Mekkakia and A. Kies: Towards a New Approach in Available Bandwidth Measures on Mobile Ad Hoc Networks, Acta Polytechnica Hungarica, Vol. 8, 2011, pp. 133-148

[3] Gy. Mester: Intelligent Mobile Robot Motion Control in Unstructured Environments, Acta Polytechnica Hungarica, Vol. 7, 2010, pp. 153-165

[4] K. Romer and F. Mattern: The Design Space of Wireless Sensor Networks, IEEE Journal on Wireless Communications, Vol. 11, 2004, pp. 54-61

[5] L. X. Guo, Y. M. Zhang and L. P. Zhao: Motion Navigation and Fuzzy Control of Mobile Robots in Wireless Sensor Networks, Sensor Letters, Vol. 9, 2011, pp. 2000-2005

[6] V. C. Gungor and G. P. Hancke: Industrial Wireless Sensor Networks: Challenges, Design Principles, and Technical Approaches, IEEE Transactions on Industrial Electronics, Vol. 56, 2009, pp. 4258-4265

[7] H. B. Lee, L. J. Park, S. W. Park, T. Y. Chung and J. H. Moon: Interactive Remote Control of Legacy Home Appliances through a Virtually Wired Sensor Network, IEEE Transactions on Consumer Electronics, Vol. 56, 2010, pp. 2241-2248

[8] A. Schoofs, G. M. P. O'Hare and A. G. Ruzzelli: Debugging Low-Power and Lossy Wireless Networks: a Survey, IEEE Communications Surveys and Tutorials, Vol. 14, 2012, pp. 311-321

[9] H. A. Tanaka, H. Nakao and K. Shinohara: Self-Organizing Timing Allocation Mechanism in Distributed Wireless Sensor Networks, IEICE Electronics Express, Vol. 6, 2009, pp. 1562-1568

[10] Y. J. Mon, C. M. Lin and I. J. Rudas: Wireless sensor Network (WSN) Control for Indoor Temperature Monitoring, Acta Polytechnica Hungarica, Vol. 9, No. 6, 2012, pp. 17-28

[11] J. S. R. Jang: ANFIS: Adaptive-Network-based Fuzzy Inference System, IEEE Transactions on Systems, Man and Cybernetics*,* Vol. 23, 1993, pp. 665-685

[12] Y. J. Mon: Airbag Controller Designed by Adaptive-Network-based Fuzzy Inference System (ANFIS), Fuzzy Sets and Systems*,* Vol. 158, 2007, pp. 2706-2714

[13] S. Kurnaz and O. Çetin: Autonomous Navigation and Landing Tasks for Fixed Wing Small Unmanned Aerial Vehicles, Acta Polytechnica Hungarica, Vol. 7, No. 1, 2010, pp. 87-102

[14] A. Azadeh, M. Saberi, V. Nadimi, M. Iman and A. Behrooznia: An Integrated Intelligent Neuro-Fuzzy Algorithm for Long-Term Electricity Consumption: Cases of Selected EU Countries, Acta Polytechnica Hungarica, Vol. 7, No. 4, 2010, pp. 71-90

[15] *MATLAB*™ fuzzy toolbox user guide (www.mathworks.com)

[16] Jennic Board API Reference Manual (JN-RM-2003), Jennic Inc., 2007 (www.jennic.com)

[17] Jennic 802.15.4 Stack API Reference Manual (JN-RM-2002), Jennic Inc., 2007 (www.jennic.com)

On Finding Better Wavelet Basis for Bearing Fault Detection

Lajos Tóth

Department of Electrical and Electronic Engineering
University of Miskolc
H-3515 Miskolc-Egyetemváros, Hungary
e-mail: elklll@uni-miskolc.hu

Tibor Tóth

Department of Information Engineering
University of Miskolc
H-3515 Miskolc-Egyetemváros, Hungary
e-mail: toth@ait.iit.uni-miskolc.hu

Abstract: This paper considers the comparision of the Meyer and Morlet wavelet for bearing fault diagnosis. We created a wavelet based upon a transient vibration signal model established for signals generated in deep-groove ball bearings with pitting (spalling) formulation on their inner race. The wavelet creation used the sub-optimal algorithm devised by Chapa and Rao that matches a Meyer wavelet to a band limited signal in two steps. We tested the applicability of the matched wavelet for identifying this kind of bearing failure. The Morlet wavelet was used as a benchmark for evaluating the performance of the matched wavelet since many publications show its successful application. It was shown that for analysing exponentially or near-exponentially damped vibration responses like the vibration produced by spalling on the inner race of a deep-groove ball bearing, the Morlet wavelet is a reasonable choice and gives better results than the Meyer wavelet.

Keywords: Wavelet analysis; bearing vibration analysis; wavelet matching; condition monitoring

1 Introduction

It is known that the *Discrete Fourier Transform* (DFT) is most suitable for testing finite-energy, periodic, time-discrete quantities. The reason why the DFT is still used effectively for the vibration analysis of bearings is that most of the complex vibrations show periodicity in time. This periodicity is closely related to the geometry and rotational speed of the bearing. Thus, the vibration components can be determined with reasonable accuracy.

Wavelet Analysis is a relatively new tool that has been successfully applied in many areas of science. In recent years, several researchers have proposed the use of wavelet transform to test bearing vibration signals where the FFT was ineffective [1-3]. Some scientists recommend the use of existing wavelet basis functions, while others create new wavelet bases. The proposed methods also differ in applied analysis techniques as well. These are *Continuous Wavelet Transform* (CWT), *Discrete Wavelet Transform* (DWT), *Wavelet-packet analysis*, *Matching Pursuit,* etc.

Junsheng et al. [4] used an impulse response wavelet to analyse faults in a roller bearing with CWT. Their wavelet is simply an exponentially damped sinusoid. Jiang et al. [5] proposed a hybrid method that combines the Morlet wavelet filter and sparse code shrinkage. Kankar et al. [6] compared three machine learning techniques for bearing fault diagnosis. These methods were the support vector machine (SVM), the artificial neural network (ANN) and self-organizing maps (SOM). For feature extraction they used a Meyer and Morlet wavelet. They found that the Meyer wavelet performs better with SVM classifier. Sheen [7] effectively applied the Morlet wavelet in the envelope detection for the vibration signal and found it also useful in the defect diagnosis of bearing vibrations. The application of the complex Morlet wavelet with SVM classifier is suggested in [8] for fault diagnosis of ball bearings having localized defects on various bearing components. Liu et al. [9] suggested an automatic feature extraction algorithm for bearing fault diagnosis using a correlation filter-based matching pursuit.

In wavelet analysis the choice of a wavelet is crucial from an analysis point of view. The analysing wavelet is usually independent of the signal investigated. Since the wavelet transformation and its derived energy distributions use convolution, one can obtain the highest output from these transformations when the signal and the wavelet are similar.

Over the past decade many publications [10-13] have considered creation of a matching wavelet to a given signal. *Tewfik* et al. [10] worked out a design method that matches a wavelet to the time domain form of a signal. *Chapa* and *Rao* [11] developed an algorithm that searches for a matching wavelet in frequency domain. Their method is capable of designing Meyer wavelets that approximate the wavelet amplitude and phase spectra separately. The cost function is the minima of the *Mean Squared Error* (MSE), calculated from the amplitude spectra and group delay of the signal and the wavelet.

Our aim was to decide whether the Meyer or the Morlet wavelet is the better choice for analysing bearing failure.

We used *Chapa* and *Rao's* algorithm to create a matching wavelet to the signal model of transient vibration generated by an artificially created fault on the inner race of a deep groove ball bearing. To evaluate its performance, we used the vibration data of a pitted single row deep-groove ball bearing of type 6209. This bearing was earlier subjected to an endurance test. We calculated the scalogram of the vibration data using the newly created wavelet and the Morlet wavelet as well. Then we compared the results and found the better representation with Morlet wavelet.

2 Methods

2.1 Wavelets, Continuous Wavelet Transform and Scalogram

Wavelets [14] are functions constructed by translating and dilating a basic function called a mother wavelet *Ψ* (see Eq. (1)). The parameters *a* and *b* are called *scale* (dilation) and *translation parameters*, respectively. The wavelet is a normalised $\|\psi\| = 1$ function.

$$\psi_{b,a}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right), \quad a > 0 \tag{1}$$

For *Ψ(t)* to be a wavelet function and to recover *f(t)* from its CWT, *Ψ(t)* should satisfy some conditions. If *Ψ(t)* has a zero average, i.e.:

$$\hat{\psi}(0) = \int_{-\infty}^{+\infty} \psi(t) dt = 0 \tag{2}$$

and satisfies the admissibility condition:

$$C_{\psi} = \int_{-\infty}^{+\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty, \tag{3}$$

where C_{Ψ} is a constant that depends on the choice of wavelet, then there exists a *Continuous Wavelet Transform* (CWT) - *Inverse Wavelet Transform* (IWT) analysis-synthesis pair [15].

The CWT of a function *f(t)* is defined as

$$CWT\,\{f(b,a)\} = \langle f, \psi_{b,a} \rangle = \int_{-\infty}^{+\infty} f(t) \cdot \frac{1}{\sqrt{a}} \psi^{*}\left(\frac{t-b}{a}\right) dt. \tag{4}$$

The benefit of CWT is that by changing the scale parameter, the duration and bandwidth of wavelet are both changed, providing better time or frequency resolution, but its shape still remains the same. The scale parameter can be linear or dyadic. The CWT uses short windows at high frequencies and long windows at low frequencies.

The *scalogram* [14], defined as the squared magnitude of CWT (Eq. (5)), always has non-negative, real-valued time-frequency (scale) distribution. This transformation conserves energy. Its resolution in the time-frequency plane depends on the scale parameter.

$$SC\{f(a,b)\} = |CWT\{f(a,b)\}|^2 = \left| \int_{-\infty}^{+\infty} f(t) \cdot \frac{1}{\sqrt{a}} \psi^*\left(\frac{t-b}{a}\right) dt \right|^2 \qquad (5)$$

Wavelets can be classified as orthogonal, bi-orthogonal, semi-orthogonal or non-orthogonal wavelets.

The Morlet wavelet is a non-orthogonal wavelet given as:

$$\psi(t) = \frac{1}{\sqrt[4]{\pi}} \left(e^{j\omega_0 t} - e^{-\frac{\omega_0^2}{2}} \right) \cdot e^{-\frac{t^2}{2}}, \qquad (6)$$

where ω_0 is the centre frequency.

In the case of $\omega_0 > 5$, Eq. (6) is simplified to (7):

$$\psi(t) = \frac{1}{\sqrt[4]{\pi}} e^{j\omega_0 t} \cdot e^{-\frac{t^2}{2}}. \qquad (7)$$

Orthogonal wavelets can be constructed from polynomial spline functions or by solving for the filter coefficients such that its Fourier transform, frequency response function, satisfies orthogonality and moment conditions [14]. Such wavelets are the *Shannon*, *Meyer*, *Battle-Lemarie* and *Daubechies* compactly supported wavelets.

The Meyer wavelet is defined through the scaling function $\phi(\omega)$ as:

$$\hat{\phi}(\omega) := \begin{cases} 1 & |\omega| < \frac{2\pi}{3} \\ \cos\left[\frac{\pi}{2} \upsilon\left(\frac{3}{2\pi}|\omega| - 1\right)\right] & \frac{2\pi}{3} \le |\omega| \le \frac{4\pi}{3} \\ 0 & \textit{otherwise} \end{cases}, \qquad (8)$$

where $\upsilon(\omega)$ is a tapering function.

2.2 Matching Wavelet to a Given Signal

Chapa and *Rao* worked out a method that is capable of designing *Meyer wavelets* to match band-limited signals. Their method directly matches a wavelet to the signal. The method requires some conditions on the wavelet spectrum amplitude and phase. They define the theorem of band-limited scaling function and the band-limited wavelet as necessary and sufficient conditions for an Orthonormal Multiresolution Analysis (OMRA). They introduce a method by which a wavelet in OMRA is expressed in terms of scaling function. They also derive constraints on the structure of the wavelet phase. The matching algorithm is sub-optimal in the sense that it matches the wavelet amplitude and phase independently [11].

2.2.1 Matching Wavelet Amplitude

The starting point of amplitude matching algorithm is the discrete form of the *refinement equation* by which a scaling function can be expressed from a wavelet:

$$\left|\hat{\phi}\left(\frac{\pi k}{2^l}\right)\right|^2 = \sum_{p=0}^{l}\left|\hat{\Psi}\left(\frac{2\pi k}{2^p}\right)\right|^2 \tag{9}$$

They define the condition on the wavelet spectra $\hat{\psi}(\omega)$ to guarantee orthonormality (Eq. (11)) and an error function (Eq. (10)) as:

$$E(Y,a) = \int_{\frac{2\pi}{3}}^{\frac{8\pi}{3}} \left[W(\omega) - aY(\omega)\right]^2 d\omega, \tag{10}$$

where:

$$W(\omega) = \left|F(\omega)^2\right|$$

$$Y(\omega) = \left|\hat{\psi}(\omega)\right|^2$$

a = scaling coefficient.

It is shown in [11] that this wavelet construction is Meyer's spectrum amplitude construction exactly. The algorithm searches for the extreme value of cost function in a discrete form. Using the symmetric property of the wavelet function, their design equation can be expressed in the form:

$$\sum_{p=0}^{l}\sum_{m=-\infty}^{+\infty} Y\left(\frac{2^l}{2^p}\left(k + 2^{l+1}m\right)\right) = 1, \tag{11}$$

where:

$$2^{l-1}/3 < \left| \frac{2^l}{2^p}\left(k + 2^{l+1} m\right) \right| < 2^{l+2}/3 \tag{12}$$

1 = vector of all ones.

Equation (11) can be rewritten in matrix form as

$$\boldsymbol{A}\, Y = 1, \tag{13}$$

where:

$$\boldsymbol{A} = \left\{ \alpha_{ij} \in \{0, 1, 2\}; \quad i = 1, \ldots, L; \quad j = 1, \ldots, 2^l \right\}$$

This amplitude matching algorithm is a constrained optimization problem that can be solved by *Lagrange multipliers* [11].

The error function (Eq. (10)) can be given by

$$E = \frac{(W - aY)^T (W - aY)}{W^T W}. \tag{14}$$

Matching amplitudes are given in the form of

$$Y = \frac{1}{a} W + \boldsymbol{A}^T \left(\boldsymbol{A}\, \boldsymbol{A}^T\right)^{-1} \left(\boldsymbol{1} - \frac{1}{a} \boldsymbol{A}\, W \right), \tag{15}$$

where

$$a = \frac{\boldsymbol{1}^T \left(\boldsymbol{A}\, \boldsymbol{A}^T\right)^{-1} \boldsymbol{A}\, W}{\boldsymbol{1}^T \left(\boldsymbol{A}\, \boldsymbol{A}^T\right)^{-1} \boldsymbol{1}}. \tag{16}$$

2.2.2 Matching Wavelet Phase

The phase matching algorithm is similar to the amplitude matching since it is based on MSE criteria, but instead of signal phase, it uses group delay. The group delay of a signal is defined as the first order, negative derivative of the phase

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega}.$$

There are specific constraints on the structure of the wavelet phase (Eq. (20)), which is expressed in terms of the phase of the scaling function [11]:

$$\theta_\Psi(\omega) = -\frac{\omega}{2} - \theta_\Phi(\omega + 2\pi) + \theta_\Phi\left(\frac{\omega}{2} + \pi\right) + \theta_\Phi\left(\frac{\omega}{2}\right), \tag{17}$$

where θ_Φ is the phase of scaling function $\phi(\omega)$ and θ_Ψ is the phase of wavelet $\Psi(\omega)$.

The wavelet phase is a symmetric, 2π periodic, even function. The method models one period of the negative of the group delay, denoted by $\lambda_T(\omega)$ as a polynomial of order *R* [11]:

$$\lambda_T(\omega) = \sum_{r=0}^{\frac{R}{2}} c_r \omega^{2r} \Pi\left(\frac{\omega}{2}\right), \tag{18}$$

where

$$\lambda(\omega) = -\tau(\omega)$$

$$\Pi(\omega) = \begin{cases} 1, & -\frac{1}{2} \le \omega < \frac{1}{2} \\ 0, & otherwise \end{cases}$$

c_r = polynomial coefficients.

By replicating one period of the group delay at every 2π interval, the group delay of the wavelet is modelled as the 2π periodic polynomial of order *R* [11]:

$$\lambda(\omega) = \sum_{k=-\infty}^{+\infty} \lambda_T(\omega - 2\pi k) = \sum_{k=-\infty}^{+\infty} \sum_{r=0}^{R/2} c_r (\omega - 2\pi k)^{2r} \Pi\left(\frac{\omega - 2\pi k}{2\pi}\right) \tag{19}$$

The discrete form of Eq. (19) can be written as (20), where $\Delta\omega = \frac{2\pi}{T}$, $P = \frac{N}{T}$, *N* is the number of samples, $-\frac{N}{2} \le n < \frac{N}{2}$.

$$\lambda(n) = \sum_{r=0}^{R/2} c_r \sum_{k=-P/2}^{P/2-1} (n - kT)^{2r} \Pi\left(\frac{n-kT}{T}\right) \tag{20}$$

Using vector notation, the group delay is expressed as

$$\lambda = \boldsymbol{B} c, \tag{21}$$

where

$$\boldsymbol{b}_{n,r} = \sum_{k=-\frac{P}{2}}^{\frac{P}{2}} (n-kT)^{2r} \Pi\left(\frac{n-kT}{T}\right). \tag{22}$$

Negatives of the group delays Λ_ψ and Λ_Φ can be expressed in terms of $\lambda(\omega)$. Applying $\Gamma_\Psi(\omega) = \Lambda_\Psi + \frac{1}{2}$ substitution [11], we get

$$\Gamma_\Psi = \boldsymbol{D}_\Psi c = -\frac{1}{2}\boldsymbol{B}_{\left(\frac{q+T}{2}\right)} + \sum_{m=2}^{\infty} 2^{-m} \boldsymbol{B}_{\frac{q}{2^m}}, \tag{23}$$

$$\Gamma_\Phi = \boldsymbol{D}_\Phi c = \sum_{m=1}^{\infty} 2^{-m} \boldsymbol{B}_{\frac{q}{2^m}}. \tag{24}$$

The matching process minimizes the weighted error (Eq. (25)) between the group delay of the wavelet Γ_Ψ and the desired signal Γ_F [11]. The approximation is performed only in the pass-band [11], thus a weighting function Ω (*n*) is calculated from the result of the amplitude matching process:

$$\gamma = \sum_{n=-N/2}^{N/2-1} \left(\Omega(n)\left(\Gamma_F(n) - \Gamma_\psi(n)\right)\right)^2, \tag{25}$$

where

$$\Omega(n) = \frac{Y(n)}{\sum Y(n)}.$$

The optimal values of the polynomial coefficients can be obtained by solving

$$\nabla_c \gamma = 0. \tag{26}$$

2.2.3 Matching a Wavelet to Signal Model of Bearing Vibration

We used the above-mentioned algorithm to create a matching wavelet to the transient vibration generated by an inner race fault – a pitting or spalling formulation – in a deep groove ball bearing. It is shown in [16] that the rolling elements generate a series of amplitude modulated transient pulses when they pass over the fault. The amplitude modulation is caused by load distribution; that is, the closer the fault is located to the load zone, the higher the amplitude of the transient is. One of these impulses can be described as:

$$y(t) = A \cdot t^n \cdot e^{-C \cdot t} \cdot \sin(\omega_0 \cdot t), \quad t \in [0, \infty) \quad \omega_n = 2\pi f_n \tag{27}$$

where f_n is the n^th^ natural frequency of bearing system, C is a damping factor, A is the initial amplitude and n is an exponent influencing the rise time of the transient.

In order to create a new wavelet basis function, we used 512 samples of time domain data of a transient vibration signal (Fig. 1) described by Eq. (27) with A=68.74, n=1.851, C=6.78, and ω_0=18.85.

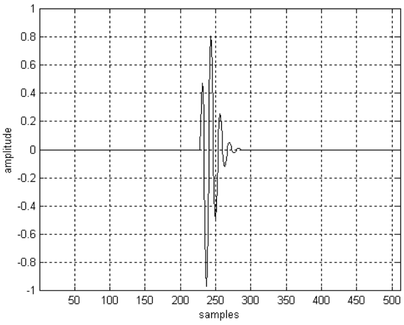


Figure 1

Transient signal model (A=68.74, n=1.851, C=6.78, ω_0=18.85)

We applied *Chapa* and *Rao's* amplitude and phase matching algorithm [11] on the transient signal model of the bearing vibration. The results are shown in Fig. 2. The new wavelet amplitude spectra match the amplitude spectra of the transient very well in the passband. The MSE of the matching is 0.011. For phase matching we used a 16^th^-degree polynomial. The matched group delay shows satisfactory characteristics in the passband. These plots are very similar to the plots introduced in [11], since the applied transient pulses are similar in both cases.

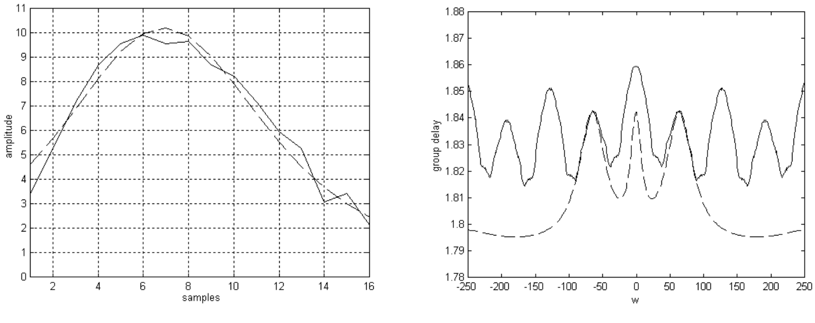


Figure 2

Matching the wavelet amplitude and the group delay in the neighbourhood of the passband (matched wavelet amplitude spectra and group delay: continuous line; amplitude spectra and group delay of the transient: dashed line)

Combining the amplitude and phase spectra together, it is possible to obtain the time domain form of wavelet function [11]. It can be seen in Fig. 3 that the new wavelet adequately fits the original transient data.

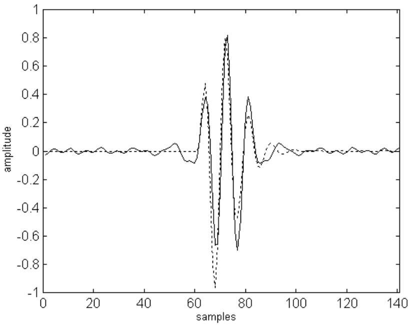


Figure 3

Transient signal and the new wavelet in the time domain (transient signal: dotted line; wavelet: continuous line)

Since the new wavelet basis cannot be given in closed form we gave the filter coefficients in Table 1. The time- and frequency behaviour of the wavelet and the impulse responses of filters are shown in Figs. 4 and 5.

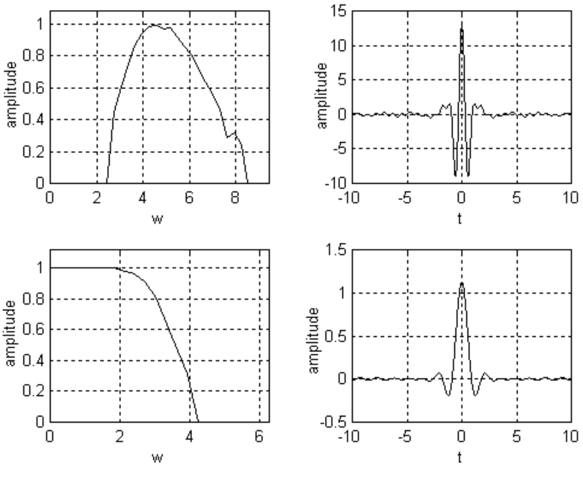


Figure 4

The new wavelet and scaling function in the frequency and time-domain

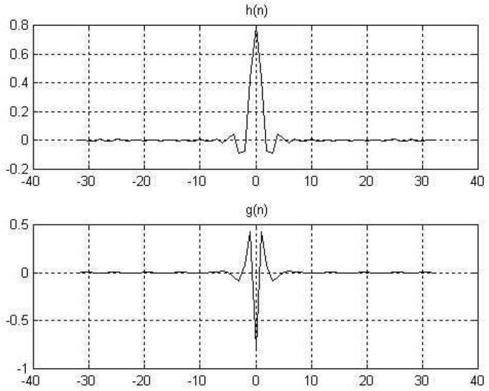


Figure 5

The impulse responses of filters corresponding to new wavelet

From Table 1, one can conclude that these *Quadrature Mirror Filters* (QMF) have no compact support. Their effective support is approximately [-10, 10]. But by limiting the filter coefficients to this range, we cannot reconstruct the original data from the wavelet coefficients completely. The fewer coefficients we use, the more error we get during reconstruction. The opposite is also true; that is, more coefficients are needed to reconstruct the original data from the wavelet coefficients. Since this wavelet was created to match the vibration signal of a bearing with a specified fault, it can be used to detect this kind of defect in a bearing.

Table 1

The new wavelet $\Psi(t)$ filter coefficients $\{h[k]\}$ and $\{g[k]\}$

k	h(n)	g(n)
± 0	0.7948	-0.7948
± 1	0.4260	0.4260
± 2	-0.0760	0.0760
± 3	-0.0872	-0.0872
± 4	0.0474	-0.0474
± 5	0.0115	0.0115
± 6	-0.0166	0.0166
± 7	0.0068	0.0068
± 8	-0.0028	0.0028
± 9	-0.0024	-0.0024
± 10	0.0069	-0.0069
± 11	-0.0045	-0.0045
± 12	-0.0012	0.0012

3 Application of Matched and Morlet Wavelet for Detection of Spalling in a Bearing

We examined the applicability of the matched and the Morlet wavelet for detecting pitting formulation on the inner race of a deep-groove ball bearing. Our aim was to detect transient pulses generated in the bearing to indicate the presence of a pitted raceway, i.e. the time of the last possible maintenance before catastrophic failure.

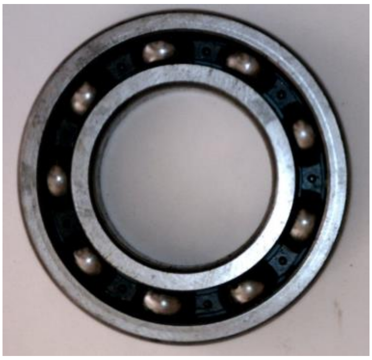


Figure 6

A single row deep-groove ball-bearing of type 6209 used as test specimen

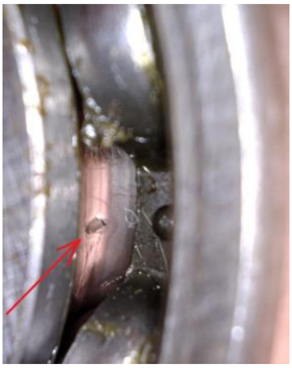


Figure 7

Pitting formulated on the inner raceway during endurance test

We used two 6209-type, single-row, deep-groove, radial ball bearings as test specimens (Fig. 6). One of them was earlier subjected to an endurance test. As a consequence of the repetitive load on the bearing elements, pitting formed on the inner ring of the bearing (Fig. 7). The other bearing was free of faults.

For the test rig we used a turning machine of E1N type. The inspected bearing was mounted on a shaft which was fixed in the chuck. We used a rod fixed in the tool post as the support. The rotating nature of the tool post made it possible to apply radial load on the outer ring of bearing, where the force was set to be perpendicular to the rotating shaft. Our primary goal was to minimize the force/vibration transmission path, since noise can come from a number of different sources. Mechanical noises can be eliminated by properly set up measurement device configuration, while electrical noises usually come from the test rig. A portion of the outer ring of bearing was machined by grinding. An accelerometer of KISTLER 8702B50 type was attached to its flat area by beeswax.

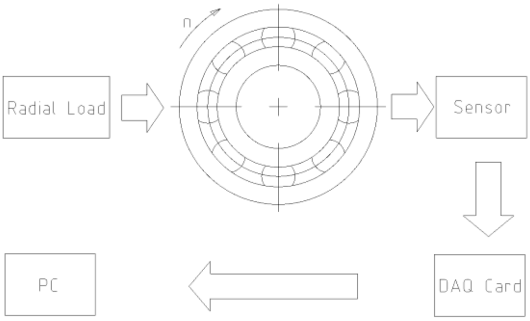


Figure 8
The measurement setup

For data acquisition (Fig. 8) we used the following devices:

- HAMEG, HM507 analog-digital oscilloscope, 100 Ms/s real-time sampling rate,
- KISTLER accelerometer 8702B50,
- KISTLER 5108 charge amplifier,
- PCI 6063E PCMCIA DAQ card, 500 ks/s sampling rate.

The DAQ card was controlled by software developed under the NI LabWindows/CVI programming environment. Validation of our software was performed using a HITACHI VG-4429 function generator and digital oscilloscope.

Sampling was performed at a constant inner ring speed of 1812 min^{-1}. This value satisfies the specifications of American ANSI [17] and German DIN [18] standards (1800 min^{-1} ±2%) concerning bearing vibration measurements. The outer ring was stationary, as it delivered radial load. The sampling frequency was set to be 20 kHz since the accelerometer's frequency range ends at 10 kHz. The gain was set to unity.

The time series data of sampled bearing vibration signal is shown in Fig. 9 and Fig. 10.

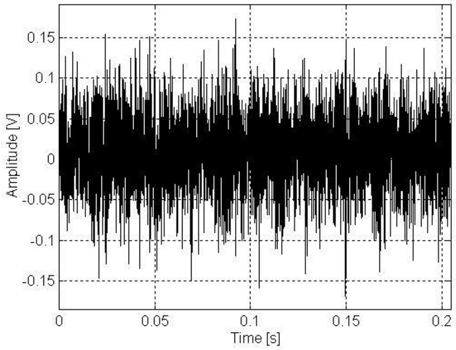


Figure 9
Vibration data of the good bearing, $f_{sampling}$=20 kHz

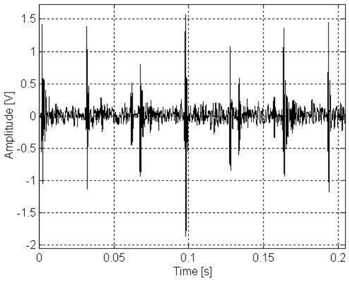


Figure 10
Vibration data of the pitted bearing, $f_{sampling}$=20 kHz

The vibration data are the results of two distinct measurements, where we could not provide exactly the same load on the bearings. Therefore one cannot make a final decision comparing the numerical values of vibration amplitudes. These figures qualitatively indicate the vibration signals. However, the difference is obviously shown. One can notice the repetitive transient pulses with an average periodicity of 30 ms. This variation in repetition time is the result of the motion of the bearing elements, which are rolling and sliding.

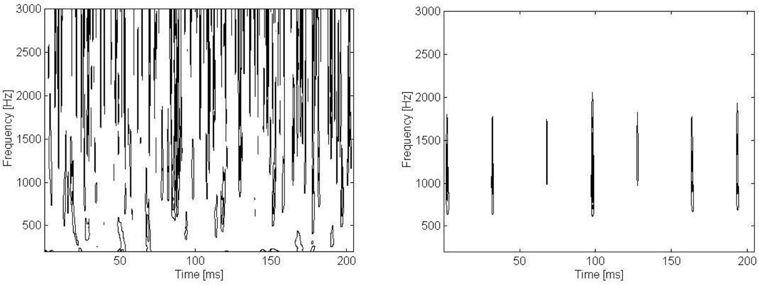


Figure 11

Time-frequency distribution of signal energy of good (left) and pitted bearing (right) using the matched wavelet

The scalogram of the good bearing clearly shows the time-frequency location of transients. These pulses do not appear at each rotation. They seem to be random signals that might come from the test rig. In contrast, periodically repeating transient pulses are clearly seen on the scalogram of the pitted bearing. Their time-period and frequency can be numerically given. The application of this method reduces the incorrect evaluation of vibration data and can be a valuable supplement to conventional condition monitoring methods.

We calculated the scalogram of the same vibration data using the Morlet wavelet as well (Fig. 12). The choice of this wavelet is obvious, since many authors report its successful application to bearing vibration analysis [5, 7, 8, 15, 19-21].

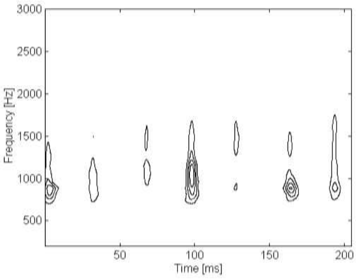


Figure 12

Time-frequency distribution of signal energy of the pitted bearing using the Morlet wavelet

Comparing the two time-frequency representations that were calculated using the matched wavelet (Fig. 11) and the Morlet wavelet (Fig. 12), we notice that the Morlet wavelet provided a more realistic result. The Morlet wavelet gave better energy concentration. The scalogram calculated by the matching wavelet provided better time localization, but its frequency localization is less accurate than that of the Morlet wavelet. This raises the question of the cause of better representation using the Morlet wavelet, since the matched wavelet was designed using the signal model of this type of failure in bearings.

4 Examination of the Wavelets and the Signal Model

The time and frequency domain plots of the matched and the Morlet wavelet are shown in Figs. 13 and 14. The wavelet amplitudes were normalized to facilitate comparison.

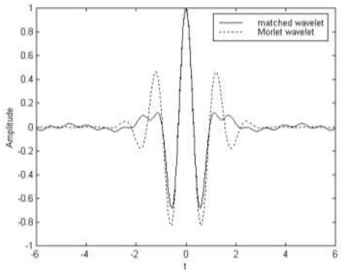


Figure 13

Time plot of the matched and the Morlet wavelet

The matched and the Morlet wavelet follow almost the same pattern between the [-1.5, 1.5] time interval, where the matched wavelet approximates the Morlet wavelet. Most of the signal energy is concentrated in this area.

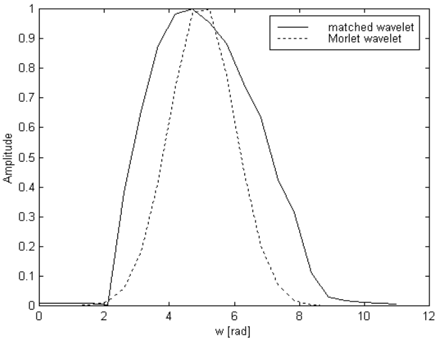


Figure 14

Amplitude spectrum of the matched and Morlet wavelet

The frequency domain form of these wavelets shows their bandpass behaviour. The bandwidth of the Morlet wavelet is narrower than that of the matched; that is, the Morlet wavelet concentrates more of the signal energy around the centre frequency. The wider bandwidth gives shorter time extent; thus the time localization ability of the matched wavelet is better. This is also clearly seen in Fig. 11. and Fig. 12.

Comparing the real part of the complex Morlet wavelet in Eq. (7) to the signal model of transient vibration produced by a pitting formulation on the inner race of a deep groove ball bearing (Eq. (27)), one can notice the similarity. These waveforms are shown in Figs. 15 and 16. The resemblance is clearly seen.

The frequency domain form of these waveforms shows their bandpass behaviour (Fig. 16). The phase characteristics of these waveforms are also similar except for the dilation caused by unwrapping their phase.

Several papers report the successful application of the Morlet wavelet in the field of bearing vibration analysis without clarifying the reason for their choice. Since the signal model of transient pulses generated by pitting on the inner raceway of a deep-groove ball bearing was very similar to the Morlet wavelet, and since the wavelet transform and its derived energy distributions give more output when the signal and the analysing wavelet are similar, we can conclude that for analysing exponentially or near-exponentially damped vibration responses like the bearing vibration signal caused by pitting formulation on the inner raceway of a deep-groove ball bearing, the Morlet wavelet is a reasonable choice that is sure to yield good results.

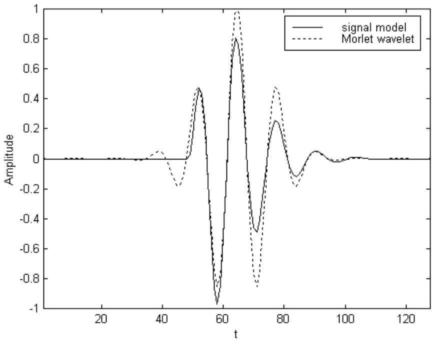


Figure 15

Signal model of bearing vibration (Eq. (30)) and the Morlet wavelet in the time domain

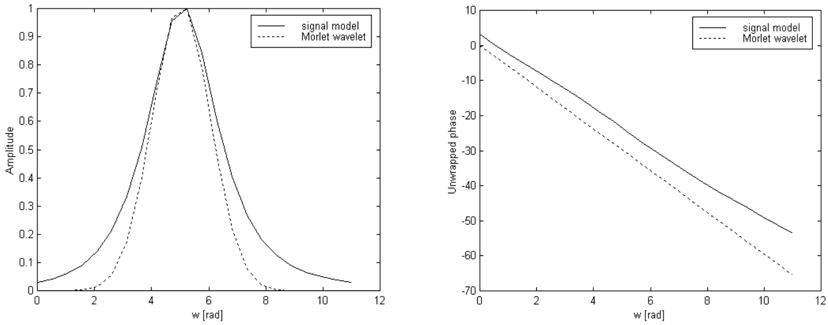


Figure 16

Amplitude and phase spectrum of the signal model and Morlet wavelet

Conclusions

This paper showed the creation of a new wavelet that matches the transient vibration response generated by pitting on the inner race of a deep-groove ball bearing. Wavelet creation is based on *Chapa and Rao's* method, where the Meyer wavelet amplitude and phase spectra are matched independently to the signal. It was shown that the new wavelet can be used for detecting transient pulses generated in a bearing. We compared the results with those calculated using the Morlet wavelet since many application reported its successful application. We found the Morlet wavelet superior to this matched wavelet in representing transient signals of bearing vibration where pitting is formulated in the inner raceway of a deep groove ball bearing. It was shown that the signal model of this kind of bearing failure is very similar to the Morlet wavelet; thus its scalogram gives a more accurate time-frequency representation than the Meyer wavelet.

Acknowledgement

This research was carried out as part of the TAMOP-4.2.1.B-10/2/KONV-2010-0001 project with support by the European Union, co-financed by the European Social Fund.

References

[1] CJ Li, J Ma, Wavelet Decomposition of Vibrations for Detection of Bearing-localized Defects. NDT & E International Vol. 30(3) (1997) pp. 143-149

[2] S. Prabhakar, A. R. Mohanty, A. S. Sekhar, Application of Discrete Wavelet Transform for Detection of Ball Bearing Race Faults, Tribology International, Vol. 35 (2002) pp. 793-800

[3] N. G. Nikolaou, A. I. Antoniadis, Rolling Element Bearing Fault Diagnosis Using Wavelet Packets, NDT & E International, Vol. 35 (2002) pp. 197-205

[4] C. Junsheng, Y. Dejie, Y. Yu., Application of an Impulse Response Wavelet to Fault Diagnosis of Rolling Bearings, Mechanical Systems and Signal Processing, Vol. 21 (2007) pp. 920-929

[5] W. He, Z. Jiang, K. Feng, Bearing Fault Detection Based on Optimal Wavelet Filter and Sparse Code Shrinkage, Elsevier - Meausrement, Vol. 42 (2009), pp. 1092-1102

[6] P. K. Kankar, S. Sharma, S. P. Harsha, Fault Diagnosis of Ball Bearings Using Continuous Wavelet Transform, Elsevier - Applied Soft Computing, Vol. 11 (2011) pp. 2300-2312

[7] Yuh-Tay Sheen, On the Study of Applying Morlet Wavelet to the Hilbert Transform for the Envelope Detection of Bearing Vibrations, Mechanical Systems and Signal Processing, Vol. 23 (2009) pp. 1518-1527

[8] P. K. Kankar, S. Sharma, S. P. Harsha, Rolling Element Bearing Fault Diagnosis Using Wavelet Transform, Elsevier - Neurocomputing, Vol. 74 (2011) pp. 1638-1645

[9] X. Liu, L. Bo, X. He, M. Veidt, Application of Correlation Matching for Automatic Bearing Fault Diagnosis, Journal of Sound and Vibration, Vol. 331 (2012) pp. 5838-5852

[10] A. H. Tewfik, D. Sinha, and P. Jorgensen, On the Optimal Choice of a Wavelet for Signal Representation, IEEE Transactions on Information Theory, Vol. 38, pp. 747-765, March 1992

[11] J. O. Chapa, R. M. Rao, Algorithms for Designing Wavelets to Match a Specified Signal, IEEE Transactions on Signal Processing, Vol. 48, No. 12, December 2000, pp. 3395-3406

[12] N. Tandon, A. Choudhury, An Analytical Model for the Prediction of the Vibration Response of Rolling Element Bearings Due to a Localized Defect, Journal of Sound and Vibration, Vol. 205 (1997) pp. 275-292

[13] A. Gupta, S. D. Joshi, Surendra Prasad, On a New Approach for Estimating Wavelet Matched to Signal. In Proceedings of Eight National Conference on Communications, Bombay, January 2002

[14] S. Mallat, A Wavelet Tour of Signal Processing, Second Edition, Academic Press, 1998

[15] J. C. Goswami, A. K. Chan, Fundamentals of Wavelets, John Wiley & Sons, Inc., 1999

[16] L. Tóth, Identification of a Transient Vibration Signal Model, microCAD 2008, International Science Conference, Section J: Electrotechnics and Electronics, pp. 83-88

[17] American National Standard ANSI/AFBMA Std 13-1970, ANSI B3.13-1970, Rolling Bearing Vibration and Noise (Methods of Measuring)

[18] Deutsches Institut für Normung DIN 5426, Laufgeräusche von Wälzlagern, Prüfverfahren

[19] R. Rubini, U. Meneghetti, Application of the Envelope and Wavelet Transform Analyses for the Diagnosis of Incipient Faults in Ball Bearings, Mechanical Systems and Signal Processing, Vol. 15 (2001) pp. 287-302

[20] S. Prabhakar, A. R. Mohanty, A. S. Sekhar, Application of Discrete Wavelet Transform for Detection of Ball Bearing Race Faults, Tribology International, Vol. 35. (2002) pp. 793-800

[21] N. G. Nikolaou, A. I. Antoniadis, Demodulation of Vibration Signals Generated by Defects in Rolling Element Bearings Using Complex Shifted Morlet Wavelets, Mechanical Systems and Signal Processing, Vol. 16 (2002) pp. 677-694

Yet Another Attempt in User Authentication

Liberios Vokorokos, Adrián Pekár, Norbert Ádám

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovakia
liberios.vokorokos@tuke.sk, adrian.pekar@tuke.sk, norbert.adam@tuke.sk

Peter Darányi

AT&T Global Network Services, Einsteinova 24, 851 01 Bratislava, Slovakia
peter.daranyi@tuke.sk

Abstract: This paper deals with the Encrypted Key Exchange (EKE) authentication method, which offers an opportunity to log on to a server and to authenticate the user itself without using a certificate or any direct transmission of the password. This makes the method an interesting alternative solution, where at the end of the authentication process generated is a key that can be further used for the needs of symmetric encryption. In the paper, an implementation of a client-server application that uses EKE authentication method is described. This application after a successful login process enables the user to transfer files in encrypted form, while the encryption key is generated at the end of the authentication process.

Keywords: authentication; Encrypted Key Exchange; Secure Password Exponential Key Exchange; shared key (secret); password; symmetric key

1 Introduction

The expert community has dealt with the issue of user authentication during login on to a server since the inception of the Internet [21]. The main goal is safe authentication using some of the available methods, such as a password or certificate. These methods must be durable to attacks whose purpose is to obtain passwords or other sensitive data. With this information (user name, password, other data, etc.) the attacker could pretend to be an authorized user and log in on a server. This paper describes one of the newer methods, the Encrypted Key Exchange [2], which combines symmetric and asymmetric cryptography. This fact led us to implement this method in an application allowing the encrypted transmission of files. At present, this method is not commonly available in any commercial application, which also contributed to the attraction of this topic.

The first chapter contains a brief description of authentication protocols (methods) using a password and their comparison. The second chapter includes a draft of the implemented protocol. In the third chapter the implementation of the program with a justification for the used implementation techniques is described. The last chapter is devoted to the verification of outputs and to evaluation of solutions.

2 Authentication Protocols

Cryptographic authentication often relies on ownership of the key authenticated by a party. Such a key usually has a length from approximately hundred bits to several thousand bits, depending on the used algorithm and the desired security level. Experience showed that people have difficulties remembering passwords having seven or eight characters. When all uppercase and lowercase letters and digits from *0* to *9* are used, a random eight-character password represents less than *48* bits of randomness. Therefore, we can conclude that even a short random key for cryptographic algorithms cannot be reliably memorized by people. Cryptographic keys are frequently stored in secure memory in computers or special equipment, such as cryptographic servers or smart cards. However, there are situations where this form of custody is inconvenient or expensive [12], [18]. For this reason, the capability to establish a secure connection that relies on short passwords easily memorized by people is desired.

3 Comparison of Selected Authentication Methods

Protocols for the creation of the key are designed to be safe in situations where participants share their password with only small entropy. At first glance it may seem impossible to achieve a key using only a short password that would not be vulnerable to brute force attack (a progressive scan of all the possibilities) to find the password. This is probably the reason why the first protocols based on passwords (password-based protocols) appeared only in *1989*. Such first protocols used the additional assumption that the client knows the server's public key without sharing password with the server. Later, Bellovin and Merritt presented a class of protocols that had this additional assumption implemented [1].

The idea of Bellovin and Merritt, the Encrypted Key Exchange (EKE) protocol [1], is that the initiator of the protocol chooses a single-shot public key and uses a shared password to encrypt the key. The respondent decrypts the public key and uses it to send a session (relational) key safely back to the initiator. Assuming that the public keys are always random strings, an attacker who is trying to sequentially test all the passwords cannot distinguish which single-shot public key was used. Further, even if the correct public key is found, it cannot be used to discover the session (relational) key because from the public key it is impossible to obtain the private key. There are many variations of the Bellovin Merritt EKE protocol as well as many alternative protocols [2], [14], [16]. Recently, the protocol was also expanded with proven security features. The original EKE protocol does not specify the used encryption algorithm, which is how to convert the password into the required key.

3.1 The Original Bellovin and Merritt's EKE Protocol

EKE is closely related to the Diffie-Hellman key agreement, and its basic idea is to transfer transient public keys, which were encrypted by a password as a shared key. Only parties knowing the password are able to complete the transfer. Parameters π (shared password) and L (security parameter) are the shared information. As in the basic Diffie-Hellman [4] exchange, the shared key (secret) is $Z_{AB} = g^{r_A r_B}$, but the algorithm for obtaining the session key K_{AB} from Z_{AB} is not specified. The protocol requires two exponentiations by both parties, which is the same as in case of the basic Diffie-Hellman [4] key exchange.

In this exchange, as described in Protocol 1, first both sides agree on large prime numbers, p, n, and an element, g $(2 \leq g \leq p - 2)$, that generates a subgroup of large order (public parameters). A chooses a random number, r_A (public key), generates $g^{r_A} = t_A$, encrypts it with π and sends to B. Sharing the password, π, B decrypts the message to obtain the shared key (secret), generates $g^{r_B} = t_B$, also generates another random number, $n_B \in_R \{1, \dots, 2_L\}$, for the session key, encrypts them and sends to A.

A		B
$r_A \in_R \mathbb{Z}_p$		
$t_A = g^{r_A}$	$\xrightarrow{A,\{t_A\}_\pi}$	$r_B \in_R \mathbb{Z}_p$
		$t_B = g^{r_B}$
		$Z_{AB} = t_A^{r_B}$
		$n_B \in_R \{1, \dots, 2_L\}$
$Z_{AB} = t_B^{r_A}$	$\xleftarrow{\{t_B\}_\pi,\{n_B\}_{K_{AB}}}$	
$n_A \in_R \{1, \dots, 2_L\}$		
	$\xrightarrow{\{n_A,n_B\}_{K_{AB}}}$	$verify\ n_B$
$verify\ n_A$	$\xleftarrow{\{n_A\}_{K_{AB}}}$	

Protocol 1

The original Bellovin-Merritt EKE Protocol

A decrypts the message to obtain the shared key (secret), also generates a random number for the session key and sends it to B. Assuming the final verification is successful, both parties can calculate the true session key that will be used for all future messages between A and B. [2], [14], [16]

3.2 The Secure Password Exponential Key Exchange Protocol (SPEKE)

The SPEKE protocol was designed by Jablon [7], [8]. Although SPEKE, like EKE, is based on the Diffie-Hellman exchange, the main difference between them is in the password used to determine the base that is used in Diffie-Hellman exchange [4]. Jablon introduced the basic and extended version of the SPEKE protocol.

Prime p and group G are chosen in a way that $q = \frac{(p-1)}{2}$ will become a prime, and G will have a degree of q. Password, π,is considered to be a number from $\mathbb{Z} *_p$ and then $P = \pi^2$ will surely lie in group G and has a degree of q (assuming that π is not equal to *1*, *-1* or *0*). Value P is used as a generator of group G. [2]

Except this special way of defining the generator of the group, the protocol is exactly the same as the basic Diffie-Hellman key exchange with key confirmation [4]. The shared key (secret) is therefore $Z_{AB} = P^{r_A r_B}$.

The original version of SPEKE has no evidence of safety, but later MacKenzie [11] introduced the proof of a slightly modified version. These changes are:

- P is defined to include the identities of A and B: $P = (H(A; B; \pi))^2$
- The hashes used to form the session key include the identities of A and B, the exchanged messages, $t_A$and t_B, the password, π, as well as the shared key (secret), Z_{AB}: $K_{AB} = (A; B; t_A; t_B; \pi; Z_{AB})$
- The exponents are chosen randomly: $r_A;\ r_B \in_R \mathbb{Z}_q$

The expanded version of the protocol, with the V_a and V_b verifiers defined in the version proposed by MacKenzie [11], is described in Protocol 2.

The SPEKE protocol consists of two stages. The first one uses Diffie-Hellman [4] to establish the shared secret (key) Z_{AB}. Where SPEKE differs from DHEKE is that, instead of the commonly used fixed primitive base g, it converts with function P the password π into a base for exponentiation. The rest of the first stage is pure Diffie-Hellman, where parties A and B start out by choosing two random numbers, r_A and r_B. A computes by function P message t_A and sends it to B. B computes by function P message t_B and sends it to A. B also computes the shared key Z_{AB}. A also computes the shared key Z_{AB}. [7]

In the second stage, both A and B confirm the knowledge of Z_{AB} before using it as session key K_{AB}. B sends with message t_B a proof of Z_{AB} (message V_B) – which is obtained by a strong one-way hash function H_1 – to A. A verifies that V_B is correct and sends its proof of Z_{AB} (message V_A) – which is obtained by a strong one-way hash function H_2 – to B. B verifies V_A is correct. Assuming the verifications are successful, the protocol is complete. [7]

A		B
$r_A \in_R \{1, \dots, 2_L\}$ $t_A = P^{r_A}$		
	$\overset{t_A}{\rightarrow}$	
		$r_B \in_R \{1, \dots, 2_L\}$ $Z_{AB} = t_A^{r_B}$ $check\ Z_{AB} \notin \{-1, 0, 1\}$ $t_B = P^{r_B}$ $V_B = H_1(t_A; t_B; Z_{AB}; \pi)$
	$\overset{t_B, V_B}{\longleftarrow}$	
$Z_{AB} = t_B^{r_A}$ $check\ Z_{AB} \notin \{-1, 0, 1\}$ $check\ V_B$ $V_A = H_2(t_A; t_B; Z_{AB}; \pi)$		
	$\overset{V_A}{\rightarrow}$	
$K_{AB} = H_3(Z_{AB})$		$check\ V_A$ $K_{AB} = H_3(Z_{AB})$

Protocol 2

Secure Password Exponential Key Exchange Protocol

In this case, shared information are as follows: subgroup G of group $\mathbb{Z} *_p$ degree $q = \frac{(p-1)}{2}$, where p and q are primes; derived password $P = \pi^2$, where π is interpreted as an element of $\mathbb{Z} *_p$; security parameter L.

For security reasons of the implemented version, and thus the application, we decided to implement the above-mentioned version of the SPEKE protocol.

4 Draft of the Implementation of Authentication Procedures

Based on [17], the SPEKE protocol was implemented in a program for file transmission. In this way, both possibilities offered by the protocol – authentication and password generation, which will be subsequently used to encrypt the transmitted files – will be used. The application is divided into two parts: Into the Secure File Share (SFS) Server and Client programs.

SFS Server is a server program and SFS Client is a client program in the client-server architecture. After start the server loads the basic settings from the settings file (if it exists), which is saved in the directory where the program runs. Otherwise, an error message appears. Next, the servers search the database for data on previously added users. Then the servers begin to listen on a specified port for user log in requests. Like the server, after the client starts, it loads the basic settings from a settings file (if it exists), and then it tries to connect to the specified IP addresses and port. After successful connection establishment and client login, on the basis of the exchanged messages both parties generate a shared key. From that moment on, until an explicit or automatic logging off, the client can transfer files located in the shared directory of the server. Files transferred over the network are encrypted.

For the implementation of cryptographic protocols, the Java programming language was chosen as it supports very large numbers (several hundred bits), hash functions and cryptographic protocols. In addition, for the widest possible usage of the resulting program, Java is not platform (PC, Macintosh, etc.) or operating system (Microsoft Windows, Linux, Mac OS X, etc.) dependent. Since the protocol is implemented for the users logging in to the server, for the user management a relational database is used, with which, on the basis of the above-mentioned criteria, Java can communicate [15]. The database server is located on the same computer as the SFS Server. The database contains only one single table with fields of all necessary data about the users. For each user registered, there are an identification number, name, password, and a flag indicating whether the user has access to the server. Field ID is the primary key table. To avoid adding more records with the same user, name entry is set as unique. For security reasons, the file transfer was only implemented from the server to the client, which will prevent a possible leakage of the key, followed by logging in of the attacker to the server and transfer potentially dangerous files to the server.

As the symmetric cryptography protocol used to encrypt the transmitted file, Advanced Encryption Standard (AES) protocol was used. AES encrypts by blocks in length of *128* bits and supports key in lengths of *128*, *192*, and *256* bits. After successful logging in using the SPEKE protocol, the client and server agree on the key that can be used for symmetric cryptography. In order to generate a sufficiently strong key and also to ensure security of the logging in to the server, primes of at least *500* bits will be used. The agreed key length has approximately the same length as the length of primes used in the protocol, which is certainly greater than the length of keys supported by AES protocol. For this reason the key for AES will be the first *128*, *192*, or *256* bits from the agreed key. In the Protocol, a hash function is also used. In order to ensure the highest security, the SHA-*512* function will be used.

5 Communication Principles

The complete authentication process and activities performed by the individual sides are described in Figure 1. After the initial initialization, the application starts to listen on a port. If someone connects to this port, another thread starts. At the beginning, the tread waits until the connecting side sends its user name. Then it sequentially checks the following things to see whether the user can log in:

- First it checks whether the user is logged in with that user name.
    - If the user is logged in the application sends an error message.
    - If not, it checks if the user name is registered in the database.
- Next, if the user is not registered it sends an error message. If yes, the application checks whether the user has not been banned on the server. When the user has passed the final check he can proceed with logging in.

The next procedure of authentication and generation of a common symmetric key is in conformity to the described SPEKE protocol in the previous chapter. If a problem occurs during the client authentication, the server immediately terminates the connection.

After successful logging in, the server sends the structure of the shared folder to the client. Then, until the end of connection, the server waits for a request from the client to resend the folder structure or for file transfer. Connection can be terminated in three ways: the client logs out, the server terminates an inactive connection after a timeout, or disconnection occurs due to network failure.

Upon termination of the connection the thread ends. In order to facilitate the work, the application was extended with a graphical user interface. On the client side, all communication (logging in, reloading the shared directory tree of the server, file transfer) with the server is performed in separate threads. With this method it is possible to save or load client settings, which consist of an IP address, server port and a local directory in which the files are downloaded.

After successful login, the structure of the shared directory tree of the server is accepted. When the user wants to transfer a file, a separate thread will start and as parameters the data obtained during logging in process are used. That is the key by which the file transfer will be encrypted. In case the connection is not successful, the application displays an error message.

During file transmission, for each file the server first sends a message that indicates the client wants to transfer a file. Pieces of the file are always transmitted in a fixed size (by default *1024* B). Before file transmission, a check is performed as to whether the file exists in the destination directory. If so, the user chooses whether to overwrite it. The user can terminate file transfer at any time.

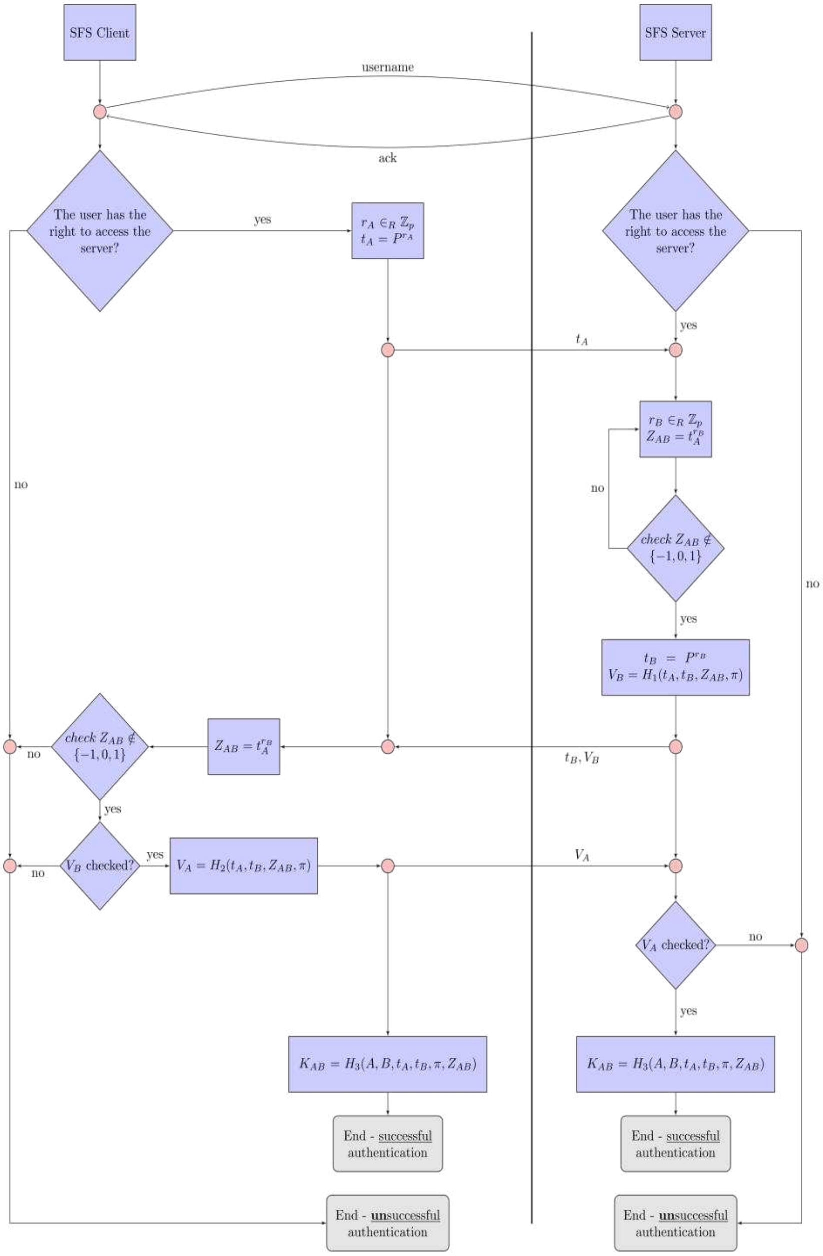

SFS Client
SFS Server
username
ack
The user has the right to access the server?
yes
no
$r_A \in_R \mathbb{Z}_p$
$t_A = P^{r_A}$
The user has the right to access the server?
yes
no
t_A
$r_B \in_R \mathbb{Z}_p$
$Z_{AB} = t_A^{r_B}$
no
check $Z_{AB} \notin \{-1,0,1\}$
yes
$t_B = P^{r_B}$
$V_B = H_1(t_A, t_B, Z_{AB}, \pi)$
t_B, V_B
$Z_{AB} = t_A^{r_B}$
check $Z_{AB} \notin \{-1,0,1\}$
no
yes
V_B checked?
yes
no
$V_A = H_2(t_A, t_B, Z_{AB}, \pi)$
V_A
V_A checked?
no
yes
$K_{AB} = H_3(A, B, t_A, t_B, \pi, Z_{AB})$
$K_{AB} = H_3(A, B, t_A, t_B, \pi, Z_{AB})$
End - successful authentication
End - successful authentication
End - unsuccessful authentication
End - unsuccessful authentication

Figure 1
Authentication process

6 Implementation of the Authentication Procedures

Due to the requirements mentioned in the draft section of this paper, we decided to use the Java programming language. This language contains the Java Cryptographic Extension (JCE) extension, by which it is possible to use common cryptographic functions. The Java Virtual Machine (JVM) allows platform independence.

Complying with [22], we chose MySQL as the database server. However, the structure of the database tables is not difficult for administering except that for the command line is also possible to use GUI tools such as MySQL Administrator. Communication with the database server consists of several steps that define the JDBC process.

Communication with the client consists of several parts. First, the message exchange happens in conformity with the SPEKE protocol described in chapter two. During this exchange, the client authenticates itself and a symmetric cryptography key is also generated. After successful completion of this phase of the program until the end of the connection, incoming requests from the client are awaited. The client can request a file transfer or retransmission of the shared directory tree structure of the server. However, if the client does not send the request within a specific amount of time (*15* minutes by default), the connection is automatically terminated. During the process of authentication and symmetric key generation, the *BigInteger* variable type is used. The problem with the variables of the common Math class is that it works with *ints* and *doubles* types. These types can hold only finite numbers and have limited precision. The *BigInteger* class can hold arbitrarily large numbers.

At first, two primes, p and q, are required to satisfy the condition $q = \frac{(p-1)}{2}$. For safety reasons it was decided to use a prime length of *500* bits. Since sending unencrypted primes is a security risk and since the generation of two large primes that meet a given condition may last for an indefinite but certainly long time, these primes are statically assigned and are to be generated by generators of cryptographically strong pseudo-random numbers.

A cryptographically strong pseudo-random number corresponds at least with the test of the statistical random number generator specified in FIPS [6]. Furthermore, *SecureRandom* must give non-deterministic output that is a cryptographically strong sequence of numbers corresponding with the description in RFC 1750, Randomness Recommendations for Security [5]. Non-deterministic output that is a cryptographically strong sequence of numbers corresponds with the description in [5]. In the following, a probable prime number, p, is generated. Then, the variable q is saved as a number, which is the result of p subtracted by *1*. Then, it is divided by *2*. Next, a new number, q, is calculated. This will continue until primes p and q are generated, which satisfy the equality $q = \frac{(p-1)}{2}$.

During initialization it is necessary to establish communication channels through which the process of authentication can be performed. I/O in Java is based on the use of streams. Input data streams read data by bytes and write them by bytes on the output. All classes for working with streams are based on abstract classes *InputStream* and *OutputStream*. After creating the communication channels, user authentication will be performed, which was described in the previous chapter. The generation of the symmetric key $K_{AB} = (A; B; t_A; t_B; \pi; Z_{AB})$ consists of sub-processes that are specific to each number (variable) of the key. After authentication and generation of the symmetric key, an initial AES setup is done.

After successful authentication of the client and preparation of AES cipher, the server starts to wait for requests from the client. The following steps were described in the previous section. Before sending a file, first it is checked whether the file exists and if not the client is informed. In the next step, pieces of the transmitted file will be determined. Subsequently, the file size is sent to the client to know how many bytes to expect. Sending the file itself is performed in a cycle. After the end of file transfer, the input stream associated with the transferred file is closed. Finally, it is checked whether the number of sent bytes is equal to file size.

7 Verifying the Implemented Method

Verification of the solution was divided into three parts: verification of the functionality of the implementation of cryptographic functions; verification of the behaviour of the application during occurrence of errors; verification of overall functionality of both programs (SFS Server and Client). These verification procedures will be the subject of interest in the following sections.

The application was also subjected to performance tests, which showed no noticeable lags.

7.1 Verification of Functionality of the Implementation of Cryptographic Functions

Verifying the functionality of the implementation of the cryptographic protocols was performed in such a way that, instead of the generation of numbers, a message fingerprint fixed assigned numbers were used, which were chosen not to meet the conditions necessary to continue the process of authentication:

- change of properly generated number Z_{AB} to a number *-1*, *0*, or *1*;
- change of the used password p;
- change of the messages from which the fingerprints were generated;
- use of different symmetric keys and initialization vectors during the encryption;
- use of user name that was denied access.

Each test resulted in positive results, i.e. the programs behaved in accordance with expectations: as soon as the error was detected the authentication process and consequently the network connection was terminated.

7.2 Verification of Behavior of the Application during Occurrence of Errors

When verifying the behaviour of the programs during error occurrence, the below errors were intentionally created:

- At the start of file transfer, the transmitted file was not in the shared directory of the server;
- During transmission, the network connection was interrupted;
- The network connection was first interrupted for client inactivity, and then the client sent a request;
- Files containing the settings were not in the directory where the program was launched;
- Interrupting the connection between the SFS Server and the database.

7.3 Verification of the Overall Functionality of the Application

When verifying the overall functionality of both programs tested, we used the situations that occur in normal use of the program without errors. The test results were as follows:

- adding and dropping a user from the database, denying and guaranteeing the access, and changing the password of an existing user was always successful;
- user authorized to access was always able to log in;
- user without access was never able to log in and he was always announced about the reason of the failure;
- the file was always transmitted successfully;
- in the Graphical User Interface the state of the component always reflect the actual state of the program.

Conclusion and Future Work

In the fast developing world of the Internet, more and more data and services are being published [9]. The number of services where client authentication is required and secure transmission of data (such as eBanking) is needed are increasing [10]. For this reason, it is necessary to improve authentication and encryption mechanisms and increase their security [19], [3]. Most programs and services that support secure data transfer use the Secure Shell (SSH) or Secure Socket Layer (SSL) technologies. Programs SFS Server and SFS Client are using a new approach, authentication with agreement on the symmetric cryptography.

During the subsequent transmission of encrypted data, the Encrypted Key Exchange (EKE) – specifically the Secure Password Exponential Key Exchange (SPEKE) – variant is used.

The advantage of EKE over commonly used applications is the fact that EKE is less common, which means that the methods of attack on this scheme are not yet known, in contrast with the methods of attack on commonly used methods, which ensures less chance of a successful attack. This advantage, with extension of EKE, will certainly become more pronounced. Another advantage of EKE is the fact that certificates are not required to prove the credibility of the server or the client, which eliminates the need for certificate generation and update.

The disadvantage of EKE against SSL and SSH is its lower versatility. Any application using EKE scheme needs to directly implement it.

Our application was designed to implement the process of authentication using the Secure Password Exponential Key Exchange (SPEKE) scheme as a login mechanism on to a server with file sharing capability, which is transmitted encrypted. Applications successfully meet these requirements. During tests it was confirmed that the processes of authentication and encryption, as well as other parts of the programs, are fully functional. Both programs were developed with regard to portability to other platforms and operating systems.

Although the presented approach should seem unnecessary, actual research and development in the field of secure file downloading mechanisms are leading to the improvement of their abilities, security and flexibility. Our approach offers a prototype of a secure file download system that implements a narrowly spread encryption protocol. The fact that this protocol is not as widely spread makes the system more unique within its application domain.

In the future, the interoperability between the system and the users will be improved by the implementation of Web technologies. The advantage of this implementation to the users is that Web applications do not need any installations on the client side, so they can be run on any system with an appropriate and compatible browser. As mentioned before, in our approach, certificates are not required to prove the credibility of the server or the client. So Web technologies will make the proposed system more flexible. This will also result in the improvements in interoperability with existing systems and clients.

Future work should also be aimed at the enhancement of the security of stored passwords and reduction of the risk that an attacker would obtain the password. This could be done by storing only the password's fingerprint $P = (H(A; B; \pi))^2$ in the database, and not the entire password. Other functions by which the applications could be extended are, for example: the generation of a new key for each file transfer; more detailed information on the client side (file size); management of the shared folder (adding, removal, renaming of files) directly in SFS Server.

Another task in the future is to extend the application by the ability to recognize the count of a session for a given user. At present, more than one user can connect to the server with the same login username and password.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-0008-10 (50%). This work is also the result of the project implementation Development of the Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).

References

[1] Bellovin, S. M, Merritt, M.: Encrypted Key Exchange: Password-based Protocols Secure Against Dictionary Attacks. IEEE Symposium on Research in Security and Privacy, 1992, pp. 72-84

[2] Boyd, C., Mathuria, A.: Protocols for Authentication and Key Establishment. Springer, p. 300, 2003

[3] Fanfara, P., Danková, E., Dufala, M.: Usage of Asymmetric Encryption Algorithms to Enhance the Security of Sensitive Data in Secure Communication. SAMI 2012, Herľany, Slovakia, 2012, pp. 213-217

[4] Diffie, W., Hellman, M.: New Directions in Cryptography. IEEE Transactions on Information Theory, 1976

[5] Eastlake, D., Crocker, S., Schiller, J.: Randomness Recommendations for Security. RFC 1750, 1994

[6] Federal Information Processing Standards Publication: Security Requirements for Cryptographic Modules. Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899-8900, 2001

[7] Jablon, D.: Strong Password-Only Authenticated Key Exchange, ACM Computer Communications Review, pp. 5-26, 1996

[8] Jablon, D.: Extended Password Protocols Immune to Dictionary Attack. Proceedings of the Sixth Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET-ICE '97) pp. 248-255, 1997

[9] Jakab, F. [et al.]: Rich Media Delivery. Computer Science and Technology Research Survey (CST '2008), Košice, 2008, pp. 31-36

[10] Kopják, J., Kovács, J.: Timed Cooperative Multitask for Tiny Real-Time-embedded Systems, IEEE 10th Jubilee International Symposium on Applied Machine Intelligence and Informatics (SAMI 2012) Herl'any, Slovakia, 2012, pp. 377-382

[11] Mackenzie, P.: On the Security of the SPEKE Password-authenticated Key Exchange Protocol. Cryptology ePrint Archive, Report 2001/057, 2001

[12] Madoš, B., Baláž, A.: Data Flow Graph Mapping Techniques of Computer Architecture with Data-driven Computation Model. SAMI 2011, Slovakia - Budapešť, 2011, pp. 355-359

[13] Michalko, M.: Video Streaming in Wireless Networks Using Avismo Concept, Journal of Information, Control and Management Systems, Vol. 9, No. 2, 2011, pp. 109-117

[14] Raymond, J., stiglic, A.: Security Issues in the Diffie-Hellman Key Agreement Protocol. IEEE Trans. on Information Theory, pp. 1-17, 2000

[15] Roman, S.: Access Database Design & Programming. O'Reilly Media, 3rd edition, 2002, p. 448

[16] Schneier, B.: Applied Cryptography: Protocols, Algorithms, and Source Code in C. Second Edition, Wiley, p. 758, 1996

[17] Szabó, CS., Slodičák, V.: Software Engineering Tasks Instrumentation by Category Theory. SAMI 2011, Slovakia, 2011 pp. 195-199

[18] Tomasek, M.: Encoding Named Channels Communication by Behavioral Schemes. Acta Polytechnica Hungarica, Vol. 8, No. 2, ISSN 1785-8860, pp. 5-19, Budapest, 2011

[19] Vokorokos, L., Ádám, N., Baláž, A.: Application of Intrusion Detection Systems in Distributed Computer Systems and Dynamic Networks. CST 2008, 2008, pp. 19-24

[20] Vokorokos, L., Baláž, A., Madoš, B.: Intrusion Detection Architecture Utilizing Graphics Processors, Acta Informatica Pragensia, Vol. 1, No. 1, 2012, pp. 50-59

[21] Vokorokos, L., Kleinová, A., Látka, O.: Network Security on the Intrusion Detection System Level. INES 2006, 2006, pp. 270-275

[22] Widenius, M., Axmark, D., Dubois, P.: Mysql Reference Manual, O'Reilly & Associates, Inc. Sebastopol, CA, USA, 2002

Analysis of Linear Interpolation of Fuzzy Sets with Entropy-based Distances

László Kovács[1] and Joel Ratsaby[2]

[1] Department of Information Technology, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary, kovacs@iit.uni-miskolc.hu

[2] Department of Electrical and Electronics Engineering, Ariel University Center of Samaria, Ariel 40700, Israel, ratsaby@ariel.ac.il

Abstract: An interpolation of fuzzy sets is an important method in development of efficient fuzzy rule systems. An important property of the interpolated set is the distance minimum property. As can be seen, the validity of this property depends on the applied distance metric. The authors analyse the distance relationship among the base and generated fuzzy sets in the case of KH linear interpolation. The paper presents new properties among the entropy-based distances and proposes an appropriate method for distance optimum interpolation.

Keywords: fuzzy interpolation; descriptive complexity; entropy; distance metric

1 Introduction

Interpolation is a widely used method to determine the values of a target function *f()* at a position *x* in a real interval [*a,b*], where *f(a)* and *f(b)* are given but *f(x)* is not known. In a more general approach, the method can be extended for an arbitrary domain *D* with $a_1, a_2, \ldots, a_n, x \in D$ to determine *f(x)* from *f*(a_1),..., *f*(a_n). Our investigation focuses on set *D* of fuzzy sets. The notion of a fuzzy set was introduced by [4]. It is a class of objects with continuous values of membership and hence extends the classical definition of a set (to distinguish it from a fuzzy set we refer to it as a crisp set). Formally, a fuzzy set is a pair *(E, m)* where *E* is a set of objects and *m* is a membership function *m : E* → [0, 1]. Fuzzy set theory can be used in a wide range of domains in which information is incomplete or imprecise, such as pattern recognition and decision theory [2] [3].

In the area of fuzzy rule interpolation (FRI) [7], the goal is to generate new fuzzy rules from existing rules. An important component of FRI is the generation of antecedent and consequent fuzzy sets using a Fuzzy Set Interpolation (FSI) method. In the most widely used approaches, *f(x)* is generated as a weighted sum of *f(a$_i$)* where the weight value depends on the distance between *x* and *a$_i$*: In the case of linear interpolation, the sum of weights is equal to 1:

$$f(x) = \sum_{i=1}^{n} w_i f(a_i), \quad \sum_{i=1}^{n} w_i = 1 .$$

The KH method developed by Kóczy and Hirota [8] uses linear interpolation as a standard FSI method. The position of the generated fuzzy set *B** is calculated with the formula

$$B^*_\alpha = \frac{\sum_{i=1}^{n} \frac{1}{d(A_\alpha^*, A_{i,\alpha})} B_{i,\alpha}}{\sum_{i=1}^{n} \frac{1}{d(A^*_\alpha, A_{i,\alpha})}},$$

where *A* denotes the antecedent set and *B* is the consequence set. The symbol α denotes a *α*-cut which is defined as $H_\alpha = \{x \in E \mid m_H(x) \geq \alpha\}$ for any *H* fuzzy set with membership function *m$_H$()*. In addition to the KH method, several new approaches are available in the literature. In the modified α-cut based interpolation (MACI) [11], fuzzy sets are described with two vectors containing the left (lower) and right (upper) flanks. The improved version of MACI is called the multidimensional modified *α*-cut based interpolation [9], and it extends MACI with the fuzziness conservation technique proposed by [10]. A more detailed survey of FRI methods can be found in [7] [12], among others.

In all versions, the distance value [1] has a central role in the interpolation algorithm. A semi-metric function to measure the distance $d : D \times D \rightarrow \Re$ meets the following conditions:

$$\begin{aligned} &d(x, y) \geq 0 \\ &d(x, x) = 0 \\ &d(x, y) = d(y, x) \\ &d(x, z) + d(y, z) \geq d(x, y) \end{aligned} . \qquad (1)$$

For the Euclidean space, the most widely used metric is the Minkowski distance between two points *x* and *y* in $\Re^n$, which is defined as

$$d_r(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{1/r}, \ r \geq 1. \qquad (2)$$

For sets in Euclidean space there are several variants for the metric function. The Hausdorff distance *q()* is defined as

$$q(U,V) = \max\left\{ \sup_{v \in V} \inf_{u \in U} d_2(u,v), \sup_{u \in U} \inf_{v \in V} d_2(u,v) \right\}. \qquad (3)$$

This can be extended to fuzzy sets as follows. Let *E* be a finite set and let *Φ(E)* be the set of all fuzzy subsets of *E*. Then, for two fuzzy subsets *A*, *B* ∈ *Φ(E),* the distance in (3) can be extended to the following distance between *A* and *B*,

$$q(A,B) = \int_0^1 q(A_\alpha, B_\alpha)d\alpha.$$

A different approach is the Hamming distance for fuzzy sets. Consider two fuzzy subsets *A, B* ∈ *Φ(E)* with membership functions m_A, m_B : *E* → [0, 1]. Then (2) can be extended to the following Hamming distance,

$$d_r(A,B) = \left(\sum\nolimits_{x \in E} \left| m_A(x) - m_B(x) \right|^r \right)^{1/r}, \ r \geq 1. \tag{4}$$

The Euclidean distance has the following nice property: consider two elements *A*, *B* in the space, then for every element *C* that satisfies

$$C = \lambda \cdot A + (1-\lambda) \cdot B, \ \lambda \in [0,1]$$

the following equality holds

$$d(A,C) + d(B,C) - d(A,C) = 0, \tag{5}$$

i.e., the points of the connecting line are extreme points from the viewpoint of distance relationship. This nice property will not in general be met for other distances.

The goal of our investigation is to analyze the relationship between the linear interpolation of fuzzy sets and the distance function in the case of a specific metric, the entropy-based distance function. The analysis shows that the fuzzy sets generated by linear interpolation will not meet (5), and a different generation method should be used to fulfill this extreme condition.

In Section 2, three basic entropy-based distance definitions for fuzzy sets are presented. The first approach corresponds to a global entropy difference, the second method is based on an element-wise entropy difference and the third approach uses a descriptive complexity with symmetric difference of the corresponding membership functions. In Section 3, the property of distance optimality is investigated in KH interpolation for the different distance interpretations. It will be shown that the KH interpolation algorithm is not suitable to generate a fuzzy set lying on the distance optimum middle point between the operand fuzzy sets. To prove the existence of such an optimum fuzzy set, a generation algorithm is also presented in the section. The theoretical considerations are demonstrated with numerical examples in the paper.

2 Entropy-based Distances

Different application areas require different similarity and distance interpretations. In the case of fuzzy sets, there are basically three main aspects of similarity [5]:

- similarity of the support set in *E* (Hausdorff metric)*;*
- similarity of the values of membership functions (Hamming metric)
- similarity of the fuzziness of membership functions

In the latter, we assume a continuous *E* domain. The fuzziness of $A \in \Phi(E)$ is defined by De Luca and Termini [6] as

$$entropy(A) = \int_{-\infty}^{\infty} S(m_A(x))dx$$

where

$$S(x) = -x\lg(x) - (1-x)\lg(1-x).$$

One approach to include the fuzziness into the distance calculation is given by the following formula:

$$d_{S1}(A,B) = \sqrt{(entropy(A) - entropy(B))^2}\ . \qquad (6)$$

As the *entropy()* function maps the fuzzy sets into $\Re^+$, $d_{s1}()$ meets the requirements of a metric function. Another way is to define an element-wise difference as

$$d_{S2}(A,B) = \left(\int_{-\infty}^{\infty} \left| entropy_A(\{x\}) - entropy_B(\{x\}) \right|^2 dx \right)^{1/2} \qquad (7)$$

where

$$entropy_A(S\{x\}) = -A(x)\lg(A(x)) - (1 - A(x))\lg(1 - A(x)).$$

This approach maps the fuzzy sets into a multi-dimensional vector space, where the applied Euclidean distance is a metric; $d_{s2}()$ meets also here the requirements of a metric function.

The third approach uses the distance function that is based on a descriptive-complexity [6]. This distance uses the symmetric difference of the corresponding membership functions and is based on the following considerations. Given two fuzzy subsets $A,B \in \Phi([N])$ with membership functions $m_A(x)$, $m_B(x)$, we denote by

$$m_{A \cap B}(x) = \min\{m_A(x), m_B(x)\}$$

and

$$m_{A \cup B}(x) = \max\{m_A(x), m_B(x)\}.$$

Define by $A \,\Delta\, B = (A \cup B) \setminus (A \cap B)$ the symmetric difference between crisp sets A,B. For fuzzy sets $A, B \in \Phi([N])$ define by

$$m_{A\Delta B}(x) = m_{A\cup B}(x) - m_{A\cap B}(x) .$$

Define a sequence of Bernoulli random variables $X_A(x)$ for $x \in [N]$ taking the value 1 with respect to $m_A(x)$ and the value 0 with respect to 1 - $m_A(x)$. Define by $H(X_A(x))$ the entropy of $X_A(x)$,

$$H(X_A(x)) = -m_A(x)\log m_A(x) - (1-m_A(x))\log(1-m_A(x)) .$$

Define the random variable

$$X_{A\Delta B}(x)) = \begin{cases} 1 & w.p. \quad m_{A\Delta B}(x) \\ 0 \, w.p. & (1-m_{A\Delta B}(x) \end{cases} .$$

We define a new distance between $A, B \in \Phi([N])$ as

$$d_{S3}(A,B) = \frac{1}{N}\sum_{x=1}^{N} H(X_{A\Delta B}(x))$$

for discrete domain and

$$d_{S3}(A,B) = \int_{-\infty}^{\infty} H(X_{A\Delta B}(x))dx \tag{8}$$

for continuous domain.

In [6] we proved that the function $d_{S3}(A, B)$ is a semi-metric on $\Phi([N])$; i.e., it is non-negative, symmetric, equals zero if $A = B$, and satisfies the triangle inequality.

Note that for any $x \in [N]$ with a crisp membership value, i.e., $m_A(x)$=1, or $m_A(x)$=0, we have $m_{A\Delta\overline{A}}(x) = 1$, and hence in this case $H(X_{A\Delta\overline{A}}(x)) = 0$. This means that for a crisp set A (for all $x \in A$, $m_A(x) \in \{0,1\}$) our distance has the following property (we call this the complement-property)

$$dist(A,\overline{A}) = 0 .$$

From an information theoretic perspective, this property is expected since knowing a set A automatically means that we also know how to describe its complement. Hence, there is no additional description necessary to describe A given its complement. This is what $dist(A,\overline{A}) = 0$ means. It can be seen from the definition that the function $dist(A, B)$ may equal zero even when $A \neq B$.

As an example, consider the fuzzy sets A,B,C and the complement A' with membership functions as shown in Figure 1. Note that A and its complement are crisp sets. The distance matrix $D = [d_{i,j}]$ is shown below; the rows and columns correspond to A, B, C and A' so that for instance the element $d_{2,3} = d_{S3}(B, C) =$ 0.709.

$$D = \begin{bmatrix} 0 & 0.354 & 0.354 & 0 \\ 0.354 & 0 & 0.709 & 0.354 \\ 0.354 & 0.709 & 0 & 0.354 \\ 0 & 0.354 & 0.354 & 0 \end{bmatrix}$$

Distance matrix D

As can be seen, *C* is a translated version of *B* and they are both the same distance from *A*. This is due to $H(X_{A\Delta B}(x)) = H(X_{A\Delta C}(x+10))$. *B* and *C* are farther apart than *B* and *A*. Since $d_{S3}(A, A') = 0$ then each one of *B*, *C* is of the same distance to *A* as to *A'*

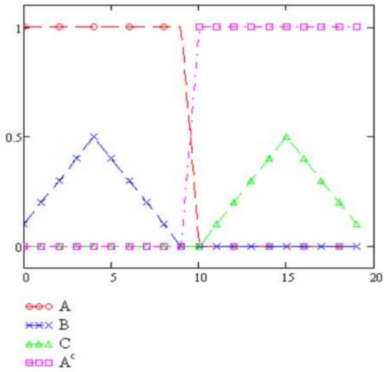


Figure 1 [6]
Fuzzy sets A,B,C and A^c

3 Distance-Optimal Interpolation Algorithm

According to (5), a linear interpolation with Euclidean metric generates elements with optimal distance. In this paper we obtained experimental results using the KH method, which was used to generate the intermediate fuzzy set *C* for given $A,B \in \Phi([N])$. In these tests, the λ value runs from 0 to 1. The test results are shown in Figure 2. In the Figure, the x-axis shows the value of λ; on the y-axis the value *ddiff(A,B,C)* = *d(A,C)* + *d(B,C)* - *d(A,B)* is given. The top (red) line is the descriptive complexity distance ($d_{S3}()$), the middle (blue) line is the element-wise entropy distance ($d_{S2}()$) and the bottom (green) line refers to the entropy-difference distance ($d_{S1}()$).

The *ddiff(A,B,C)* value indicates whether the generated *C* element is the closest element to both *A* and *B*. If *ddiff(A,B,C)* is equal to zero, the triangle inequality yields an equality and *C* lies on the line connecting *A* to *B*.

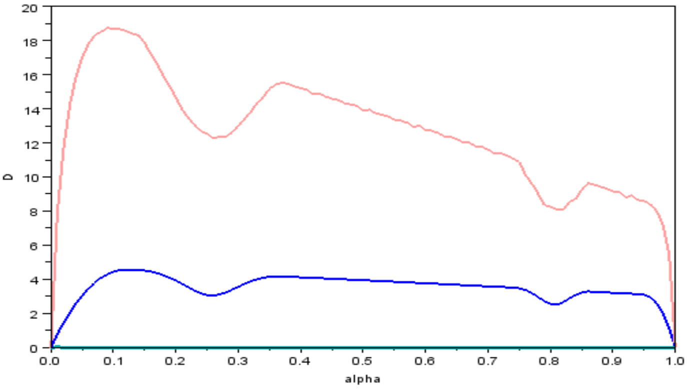


Figure 2

Distance differences for $d_{S1}()$,$d_{S2}()$ and $d_{S3}()$

Based on the test results, we conclude the following:

Property 1: For the entropy-difference distance $d_{S1}()$, for elements generated by KH interpolation, the distance difference *ddiff(A,B,C)* is equal to zero.

Proof. Let us take trapezoid membership functions with the following parameters for a set *A*:

$$\begin{aligned} A_1 &= \inf\{A_{\alpha=0}\} \\ A_2 &= \inf\{A_{\alpha=1}\} \\ A_3 &= \sup\{A_{\alpha=1}\} \\ A_4 &= \sup\{A_{\alpha=0}\} \end{aligned} .$$

where symbol $A_{\alpha=c}$ denotes the set of points with the membership function *A* equal to *c*. The *entropy(A)* differs from zero only on the intervals (A_1,A_2) and (A_3,A_4). The entropy value *entropy((A1,A2))* is calculated with

$$-\int_0^{A_2-A_1}\left(\frac{x}{A_2-A_1}\log\left(\frac{x}{A_2-A_1}\right)+(1-\frac{x}{A_2-A_1})\log(1-\frac{x}{A_2-A_1})\right)dx\,.$$

With corresponding substitutions, the integral can be transformed into the form

$$-2(A_2-A_1)\int_0^1 z\log(z)dz=-2(A_2-A_1)\left[\frac{2z^2\log(z)-z^2}{4}\right]_0^1=\frac{A_2-A_1}{2}\,.$$

Thus, the entropy value for the set A, is equal with

$$entropy(A(A_1,A_2,A_3,A_4))=\frac{(A_2-A_1)+(A_4-A_3)}{2}\,,$$

i.e., it is equal to the length of it non-crisp parts. Taking a *C* KH-interpolated set with parameter λ, the *C* will be also a trapezoid fuzzy set with the following parameters:

$$C_i = \lambda \cdot A_i + (1-\lambda) \cdot B_i .$$

It follows from the linearity that also

$$entropy(C) = \lambda \cdot entropy(A) + (1-\lambda) \cdot entropy(B)$$

holds. Thus,

$$entropy(C) \in [\min\{entrpoy(A), entropy(B)\}, \max\{entrpoy(A), entropy(B)\}]$$

and *ddiff(A,B,C)* = 0 is met.

Assuming the membership function can be approximated with a chain of linear segments, the *ddiff(A,B,C)* = 0 condition is fulfilled for fuzzy sets of arbitrary shapes. ■

Property 2. For every $A,B \in \Phi(E)$, the $d_{S3}(A,B) \geq d_{S2}(A,B)$ inequality holds.

Proof. Consider first the following inequality,

$$H(X_{A\Delta B}(x)) \geq |entropy_A(\{x\}) - entropy_B(\{x\})| . \tag{9}$$

for every $x \in E$. The inequality in (9) can be converted into the following expression:

$$K(x) = entropy_{A\Delta B}(\{x\}) - |entropy_A(\{x\}) - entropy_B(\{x\})| \geq 0 .$$

The *entropy()* function can be substituted with its definition:

$$\begin{gathered} K(x) = -|x_A - x_B| \log(|x_A - x_B|) - (1-|x_A - x_B|)\log(1-|x_A - x_B|) - \\ |-x_A \log(x_A) - (1-x_A)\log(1-x_A) + x_B \log(x_B) + (1-x_B)\log(1-x_B)| \end{gathered} . \tag{10}$$

where x_A denotes $m_A(x)$.

Let us fix x_b to a value *b* and simplify notation x_a to *x*. As (10) contains two absolute value expressions, four different subdomains should be defined:

$$\begin{gathered} R1: x < b, entropy(x) < entropy(b) \\ R2: x > b, entropy(x) < entropy(b) \\ R3: x > b, entropy(x) > entropy(b) \\ R4: x < b, entropy(x) > entropy(b) \end{gathered} .$$

In subdomain R1, formula (10) can be written as

$$\begin{gathered} K(x) = -(b-x)\log(b-x) - (1-b+x)\log(1-b+x) + \\ b\log(b) + (1-b)\log(1-b) - x\log(x) + (1-x)\log(1-x) \end{gathered} .$$

The extreme point of *K()* meets the following equation

$$\frac{\partial K}{\partial x} = -\log(b-x) + \log(1-b+x) - \log(x) + \log(1-x) = 0\,.$$

This yields in

$$\frac{(1-b+x)x}{(b-x)(1-x)} = 1$$

and

$$x = \frac{b}{2}\,.$$

In subdomain R1, the extreme points lie on the line *y* = *2x*. In a similar way, the extreme points are the following in the other subdomains:

$$\begin{aligned} &R1: y = 2x \\ &R2: y = 2x-1 \\ &R3: no\,solution \\ &R4: no\,solution \end{aligned}\,.$$

As can be easily verified, the following conditions are met:

$$\begin{aligned} &K(0) = 0 \\ &K(1) = 0 \\ &K(b) = 0 \end{aligned}\,.$$

Thus, for every $b \in [0,1]$, the K(x) function has the following function-value segments: zero, increasing, decreasing, zero, increasing, decreasing, zero. From this fact, it follows that

$$K(x) \geq 0$$

for every x and b value. Thus condition (9) is met. The measured K() values are given in Figure 3.

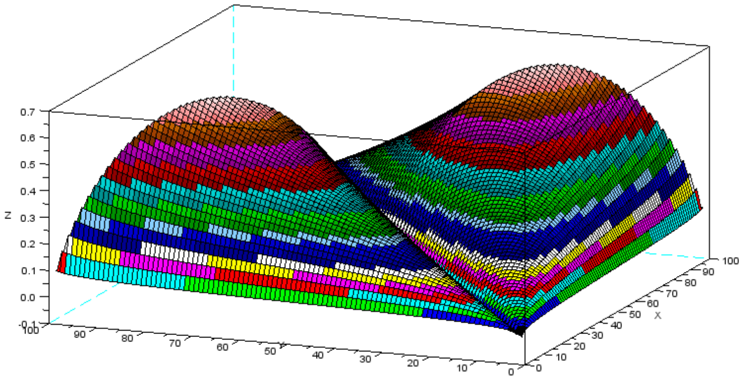


Figure 3
The K() difference function

From the fact

$$H(X_{A\Delta B}(x)) \geq \left| entropy_A(\{x\}) - entropy_B(\{x\}) \right|$$

it follows that

$$\sum H(X_{A\Delta B}(x))^2 \geq \sum \left| entropy_A(\{x\}) - entropy_B(\{x\}) \right|^2$$

and

$$\left(\sum H(X_{A\Delta B}(x)) \right)^2 \geq \sum H(X_{A\Delta B}(x))^2 \geq \sum \left| entropy_A(\{x\}) - entropy_B(\{x\}) \right|^2 .$$

Extending the expression to infinite elements, we get the expected property

$$d_{S3}(A,B) \geq d_{S2}(A,B) . \blacksquare$$

As can be seen, the KH interpolation algorithm is not suitable to generate a fuzzy set *C* lying on the middle point between *A* and *B*, i.e.

$$d_{S3}(A,C) = d_{S3}(B,C) = \frac{d_{S3}(A,B}{2} .$$

In the next step, an algorithm is presented for generating the required *C* set.

Property 3. The required C_λ set can be generated from *A*, *B* in such a way that every elements of C_λ,(*x*) is either equal to *A(x)* or to *B(x)*.

Proof. For the required element C_λ, the equation

$$ddiff(A,B,C_\lambda) = d_{S3}(A,C_\lambda) + d_{S3}(B,C_\lambda) - d_{S3}(A,B) = 0$$

should be met. It follows from definition (8) that

$$ddiff(A,B,C_\lambda) = \int_{-\infty}^{\infty} H(X_{A\Delta C_\lambda}(x)) + H(X_{B\Delta C_\lambda}(x)) - H(X_{A\Delta B}(x))dx .$$

In a similar way, as was shown in the proof of Property 2, we get

$$H(X_{A\Delta C_\lambda}(x)) + H(X_{B\Delta C_\lambda}(x)) - H(X_{A\Delta B}(x)) \geq 0, \forall x \in [0,1]$$

and

$$dd(x) = (H(X_{A\Delta C_\lambda}(x)) + (H(X_{B\Delta C_\lambda}(x)) - H(X_{A\Delta B}(x))) = 0$$

if and only if

$$\begin{array}{l} m_{C_\lambda}(x) = m_A(x) \; or \\ m_{C_\lambda}(x) = m_B(x) \end{array} .$$

If $m_A(x)$ = 1 (or = 0) then $m_C(x)$ can be equal to zero (or 1) too. The same is true for $m_B(x)$ also. ■

Figure 4 shows the $dd(x)$ value for $m_C(x) \in [0..1]$, $m_A(x) = 0.1$, $m_B(x) = 0.7$.

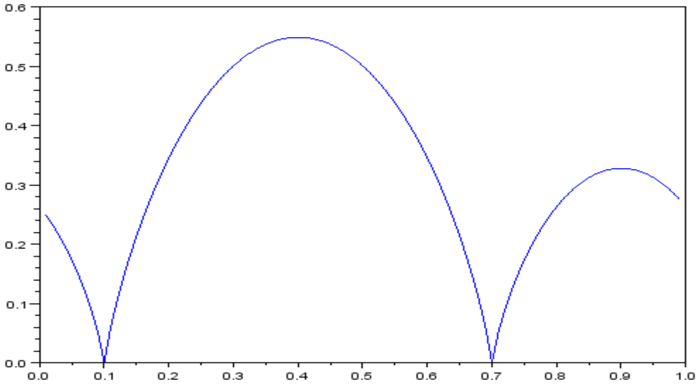


Figure 4

The dd() difference function

Based on this result, a constructive algorithm can be given to generate C_λ from the sets A and B. The algorithm assigns points to C_λ from A in a greedy way, until it reaches the required distance value:

```
Gen($\lambda$,A,B)
C = B
i = 1
while ($d_{S3}(A,C) > \lambda d_{S3}(A,B)$) {
        C =A(0..i) $\cup$ C (i+1..N)
        i++
}
```

In Figure 5, the fuzzy sets generated by KH and the proposed Gen() function are displayed. The two target trapezoid fuzzy sets *A* and *B* are shown in Figure 5a. The KH interpolated fuzzy set *C′* with λ=0.5 is given in Figure 5b in the middle in a solid blue line. The interpolated fuzzy set C" generated with Gen() is shown in Figure 5b with a thick brown line.

In the example, the following distance values can be measured:

$$d_{S3}(A,B) = 62.45$$
$$d_{S3}(A,C') = 56.20$$
$$d_{S3}(B,C') = 68.71$$
$$d_{S3}(A,C'') = 31.23$$
$$d_{S3}(B,C'') = 31.23$$

Thus, the Gen() method yields the required distance relationship for the interpolated *C* set using the descriptive complexity distance.

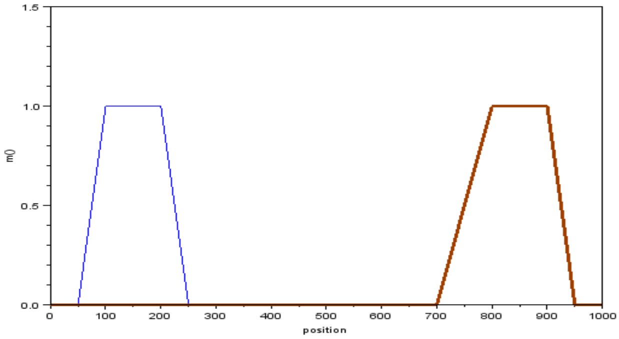


Figure 5a
The A and B fuzzy sets

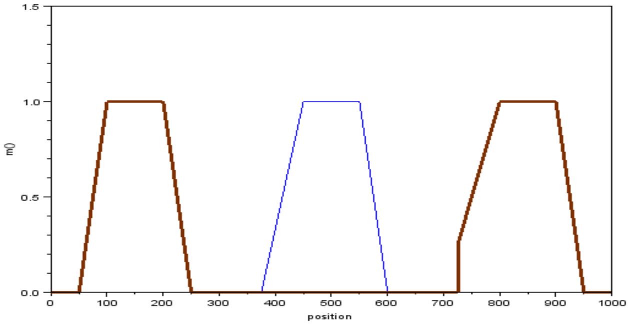


Figure 5b
The interpolated C sets

Conclusion

This paper analyzes the distance relationship among the base and generated fuzzy sets for KH linear interpolation. In the case of Euclidean distance, the usual behavior can be seen, but in the case of entropy-based distances, the new generated sets do not provide the distance optimum. The paper presents new properties among the entropy-based distances and proposes an appropriate method of distance optimum interpolation.

Acknowledgement

This research was supported by the Hungarian National Scientific Research Fund Grant OTKA K77809.

References

[1] M. Deza and E. Deza. *Encyclopedia of Distances*, Vol. 15 of Series in Computer Science. Springer-Verlag, 2009

[2] J. Ratsaby. Information Efficiency. In *Proc. of 33rd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM '07)*, Vol. LNCS 4362, pp. 475-487, 2007

[3] J. Ratsaby, Information Set-Distance, *Proc. of the 2010 Mini-Conference on Applied Theoretical Computer Science (MATCOS 2010)*, pp. 61-64, University of Primorska Press, Koper, Slovenia, 2011

[4] L. A. Zadeh. Fuzzy Sets. *Information Control*, 8:338-353, 1965

[5] R. Zwick, E. Carlstein, and D. V. Budescu. Measures of Similarity among Fuzzy Concepts: A Comparative Analysis. *International Journal of Approximate Reasoning*, 1:221-242, 1987

[6] L. Kovács, J. Ratsaby: Descriptive-Complexity-based Distance for Fuzzy Sets, CoRR abs/1012.3410: (2010)

[7] Zs. Cs. Johanyák, Sz. Kovács: A Brief Survey and Comparison on Various Interpolation-based Fuzzy Reasoning Methods, *Acta Polytechnica Hungarica*, Vol. 3, No. 1, 2006, pp. 91-105

[8] Kóczy, L. T., Hirota, K.: Rule Interpolation by α-Level Sets in Fuzzy Approximate Reasoning, In *J. BUSEFAL, Automne, URA-CNRS*, Vol. 46, Toulouse, France, 1991, pp. 115-123

[9] Wong, K. W., Gedeon, T. D., Tikk, D.: An Improved Multidimensional α-Cut-based Fuzzy Interpolation Technique, In *Proc. Int. Conf Artificial Intelligence in Science and Technology (AISAT 2000)*, Hobart, Australia, 2000, pp. 29-32

[10] Gedeon, T. D., Kóczy, L. T.: Conservation of Fuzziness in the Rule Interpolation, Intelligent Technologies, *Int. Symposium on New Trends in Control of Large Scale Systems*, Vol. 1, Herl'any, 1996, pp. 13-19

[11] Tikk, D., Baranyi, P.: Comprehensive Analysis of a New Fuzzy Rule Interpolation Method*, IEEE Trans Fuzzy Syst.*, Vol. 8, June 2000, pp. 281-296

[12] Perfiliva, I., Wrublova, M., Hodakova, P.: Fuzzy Interpolation According to Fuzzy and Classical Conditions, *Acta Polytechnica Hungarica*, Vol. 7, No. 4, 2010, pp. 39-54

Thermal Factors of Die Casting and Their Impact on the Service Life of Moulds and the Quality of Castings

Darina Matisková [(1)] **, Štefan Gašpar** [(2)] **, Ladislav Mura** [(3)]

[(1),(2)] Faculty of Manufacturing Technologies of Technical University in Košice with seat in Prešov, Department of Manufacturing Management and Department of Designing Technological Devices, Bayerova1, 080 06 Prešov, Slovakia
e-mail: darina.matiskova@tuke.sk

[(3)] J. Selye University in Komárno, Faculty of Economics, Bratislavská cesta 3322, 945 01 Komárno, Slovakia, e-mail: mural@selyeuni.sk

Abstract: This article deals with the analysis of the temperature factors of die casting and the conditions of the service life of moulds. It also describes a mechanism of the origin of internal structures and development of crystallisation grains of aluminium castings depending on the degree of undercooling. The monitored factors are very important in terms of production efficiency and the quality of the casting, which is positively reflected in the most important economic indicators of the production. In die casting, the melted metal is pressed under high temperature into a mould cavity of significantly lower temperature. The mould is then exposed to thermal strain of individual surface layers of the mould material. The speed of cooling of the castings in the metal mould increases, causing an increase of thermal gradients in the casting. Intensive heat exchange between the casting and the metal mould has a negative effect on fluidity, which increases the danger of incomplete replenishment and the occurrence of cold joints.

Keywords: thermal factors of die casting; service life of a mould; quality of a casting

1 Introduction

Die casting is a process for producing castings by die casting machines, where molten metal is injected into a permanent mould at high speed (10 - 100 m.s^{-1}) and under high pressure, Fig. 1. It is a highly productive method of die casting with low dimensions tolerance and high surface quality. The quality of the castings filled under pressure is influenced by many technological factors, the basics of which are: speed of pressing during the cycle of casting, after pressure, the temperature of cast alloy, the temperature of the filling chamber, and the temperature of a mould.

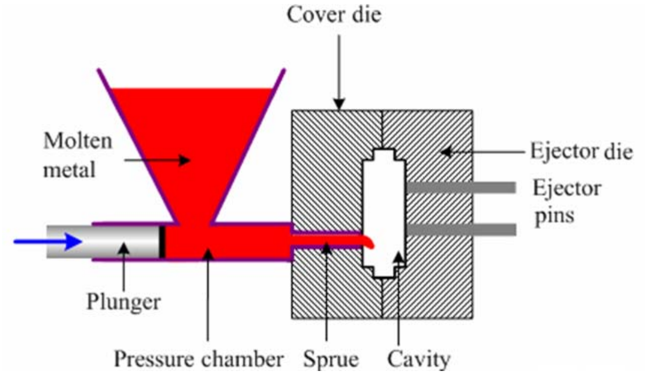


Figure 1
Scheme of die casting process

In die casting, the aforementioned input cost is divided among a large number of the products in a batch. It allows applying the solutions that secure the dimensions and shapes of the castings that are similar to the final parts with the highest possible quality of surface. Small or even no allowances for machining decrease material cost, number of operations, and overall work expenditure. This results in low weight, decreased price of the parts, high efficiency, and high competitiveness of the technology. [1, 3]

Despite recent scientific research engaged in the issue of die casting, many factors and problems related to this technology remain unexplained. It also refers to the study of impact of thermal factors of casting on the service life of the moulds and the quality of the castings.

2 Thermal Factors of Die Casting

The basic thermal factors in the die casting process are: the temperature of cast alloy, the temperature of pressurised casting chamber, and the temperature of mould.

A significant precondition for the production of high quality castings is keeping an optimum temperature of the respective parts of the mould cavity surface. This temperature depends on the temperature of the material, the quantity of the metal, the method of the cooling of the casting mould, the thermal conductivity of the mould material, and the time during which the casting remains in the mould. Casting of too hot a material into a cold mould without sufficient surface isolation by a suitable lubricating medium causes great straining of the surface layers of the casting mould material. Casting of an alloy into a mould with insufficient surface temperature results in an early fall in the alloy temperature. The castings then have cold joints (Figure 1), breaks appear on the surface of castings (Figure 2), and even good looking castings are not of necessary quality, as in the material structure large internal strains occur due to great undercooling. In some cases it is demonstrated by fine surface cracks (Figure 3). In casting the alloy into a mould with high temperature of the mould cavity surface, diffusion of the alloy into the overheated mould surface occurs. After solidification, it is demonstrated by intensive adherence of alloy to a mould and increased bubbles and porosity.

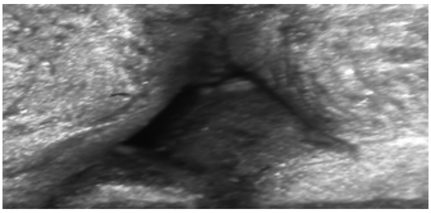


Figure 2
Cold joint

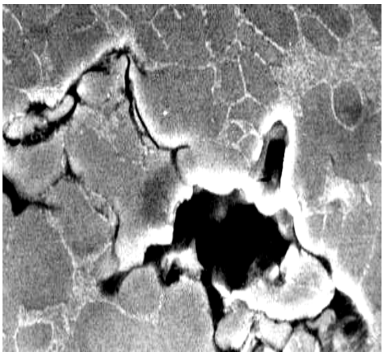


Figure 3
Internal scar

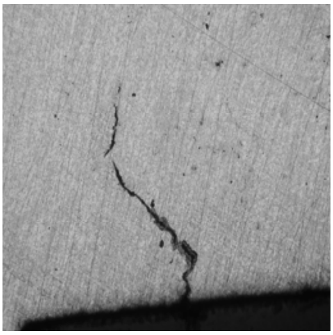


Figure 4
Surface cracks

Temperature of a cast alloy – The casting of a very warm alloy into a cold mould cavity causes serious strain on the surface layers of the mould material. The temperature of the alloy for die casting has to be higher by about 10 – 20 ºC than the initial crystallization temperature.

Temperature of a mould – the casting quality is significantly affected by the temperature regime of the mould. When the mould is cold, joints are formed in the casting and the surface quality of casting is insufficient. When the mould is very hot, the alloys are bonded to the mould and the amount of bubbles and pores grows. The temperature of the mould will be maintained constantly at approximately 1/3 of the temperature of the cast metal. This is achieved by thermoregulation of the mould.

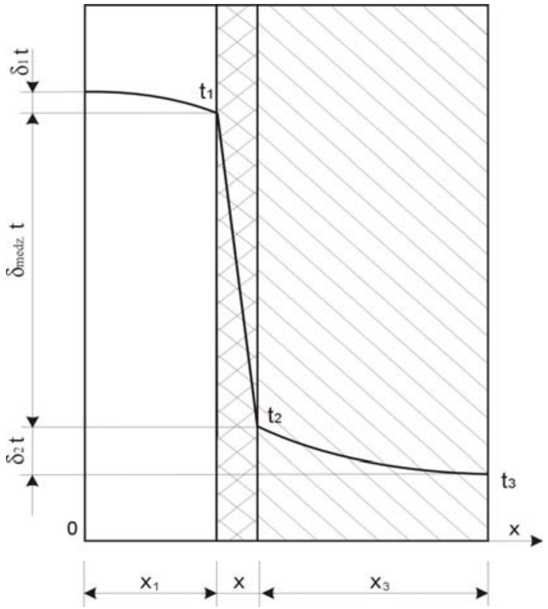


Figure 5
Temperature field of casting and mould

Temperature of a filling chamber – when the chamber is preheated before the casting process, the temperature of the cast alloy does not decrease before filling of the mould cavity.

The increased rate of cooling of castings in a metal mould in comparison with a sand mould causes an increase in temperature gradients in the casting. As a result, cure speed increases. Also the wide thickness of the gap between the casting and the metal mould, formed by a layer of the protective coating, causes temperature distribution, which is shown schematically in Figure 5. Basic thermal resistance is represented by the gap in which there is also the biggest temperature (gradient).

One of the key issues of the metal moulds heat regime is keeping the balance of one thermal cycle.

The following applies for the thermal balance per one cycle:

$$Q_1 + Q_2 - Q_3 - Q_4 = Q_1 \tag{1}$$

Where:

Q_1 - is the amount of heat required for preheating of the moulds [J],

Q_2 - is the amount of heat injected by casting [J],

Q_3 - is the amount of thermal energy transferred from the casting to the mould [J],

Q_4 - is the amount of heat minus the heat of the mould leaving the casting [J],

or:

$$Q_2 - Q_3 - Q_4 = 0 \, . \tag{2}$$

Then the amount of heat injected by the casting:

$$Q_2 = m_k\left[c_L(T_1 - T_s) + l + c_S(T_s - 20)\right], \tag{3}$$

where:

m_k - is weight of casting + 0.6 of intake weight [kg],

c_L - is the specific heat of molten iron [J.kg^{-1}.K^{-1}],

T_1 - is the temperature of the liquid metal [ºC],

T_s - is the solidus temperature [ºC],

l - is the latent heat of metal [J.kg^{-1}],

c_S - is the specific heat of solid metal [J.kg^{-1}.K^{-1}].

The amount of heat transmitted by the casting to the mould:

$$Q_3 = \alpha.F_{odl.}\left(T_1 - T_f\right)\tau_1, \tag{4}$$

where:

α - is the heat transfer coefficient [W.m^{-2}.K^{-1}],

F_{odl} - is the casting surface [m^2],

T_f - is the temperature of the mould [ºC],

τ_1 - is the cooling time [s].

The amount of heat leaving the mould with the casting:

$$Q_4 = m_k.c_s\left(T_2 - 20\right), \tag{5}$$

where:

T_2 - is the temperature of the casting taken out from the mould [ºC].

The heat balance equation is as follows:

$$m_k\left[c_L\left(T_1-T_s\right)+l+c_S\left(T_s-20\right)\right]-m_k.c_s\left(T_2-20\right)=\alpha.F_{odl.}\left(T_1-T_f\right)\tau_1 \tag{6}$$

3 Thermal Stress and Service Life of Moulds

As per Figure 6, heat from the liquid metal passes into a mould surface and further through a mould wall.

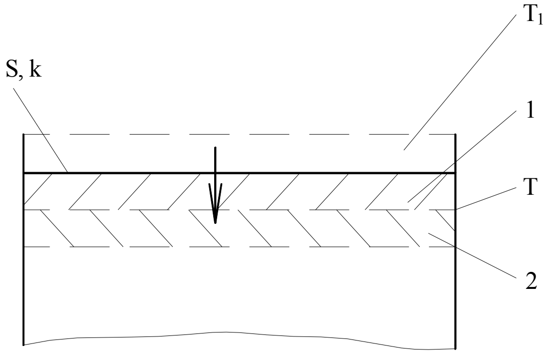


Figure 6
The surface layer 1 and the undersurface layer 2 at filling with the liquid metal

Then the heat differential *dQ* from the metal passes into the mould during the time differential *dt:*

$$dQ=\alpha S(T_1-T)dt=mcdt\,, \tag{7}$$

where:

α - is the coefficient of heat transfer between the metal and the mould surface, [W.m^{-2}K^{-1}],

S - is the contact area of the liquid metal and the mould [m^2],

T_1 - is the temperature of the liquid metal [°C],

T - is the temperature of the mould [°C],

m - is the mass of the mould surface layer [kg],

c - is the specific heat capacity of the mould material, [J.kg^{-1}K^{-1}].

By arrangement and integration of the equation (8) we will get

$$\int_0^t \frac{kS}{mc} dt = -\int_{T_1}^{T} \frac{d(T_1 - T)}{T_1 - T} \tag{8}$$

After the integration:

$$T = T_1(1 - e^{-\frac{kS}{mc}t}) \text{ [°C]} \tag{9}$$

We can write for the heat transfer from the surface layer of the temperature T_1 into the mould body of the temperature T during the time interval dt:

$$-mcdT = kS(T - T_1)dt \tag{10}$$

After arrangement and integration:

$$-\int_{T_1}^{T} \frac{d(T - T_1)}{T - T_1} = \int_o^t \frac{kS}{mc} dt \tag{11}$$

After further arrangement:

$$T = T_1 e^{-\frac{kS}{mc}t}, \text{ [°C]} \tag{12}$$

The derived courses correspond to the measured results [1, 3] in Figure7.

t - it is time [s]

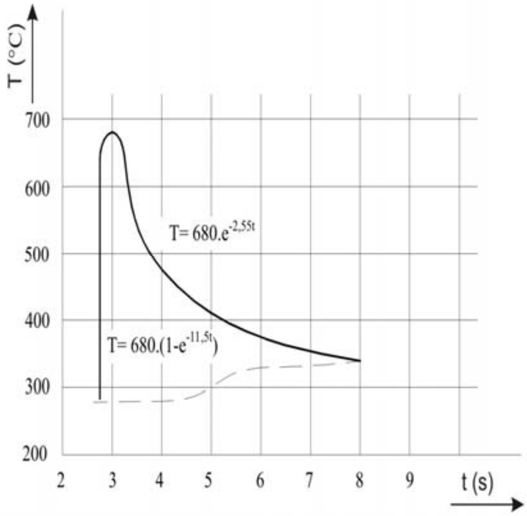


Figure 7

Course of the temperature in the experimental mould in die casting of aluminum alloy

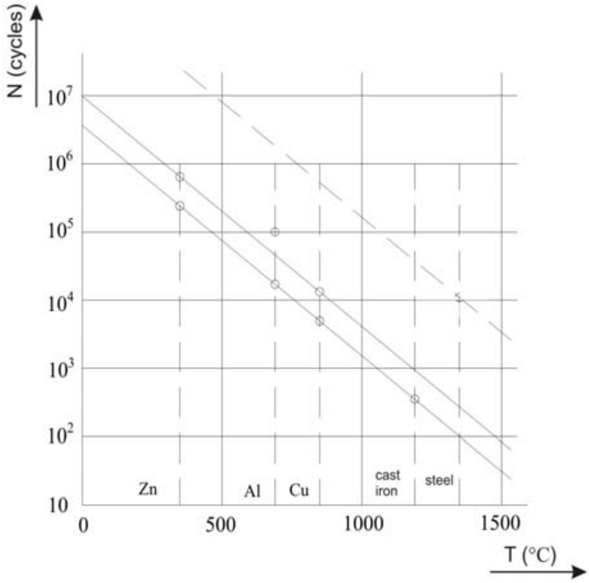


Figure 8
Dependence of the mould service life in die casting on casting temperature

To make it simple, we consider the surface layer of the mould as a total. In case the layer is loose after the liquid metal enters the mould, then dilatation occurs.

$$\delta = \alpha(T_1 - T_0) \qquad (13)$$

where:

δ - is dilatation of the mould [m],

α - is the linear coefficient of heat transfer [W.m^{-2}K^{-1}],

T_1 - is the temperature of the mould after the liquid metal enters the mould [°C],

T_0 - is the temperature of the mould before the liquid metal enters the mould [°C].

Because the surface layer is a part of the mould, it cannot dilate, and compressive stress arises in it:

$$\sigma = \frac{\delta \kappa E}{\kappa - 1} \qquad (14)$$

where:

σ - is the compressive stress of the layer [Pa],

κ - is the Poisson constant,

E - is the modulus of mould elasticity [Pa].

This compressive stress causes tensile in the layer under the surface layer. However, after removing the casting at cooling of a previously under pressure deformed surface layer the tensile stress arises in it. By repeating this process as the thermal fatigue in the tensile stress there is a danger of crack formation. After hundreds of thousands or more, cycles it really happens, and thus the mould service life ends (Figure 8).

We consider the number *N* of pressings of the liquid metal into the mould as a quasilinear function with an indirectly proportional dependence on the casting temperature *T*. Each elementary increase of the casting temperature *dT* means a lowering of the indirectly proportional cycles of the mould service life

$$-\frac{dN}{N} \tag{15}$$

Then it applies that

$$dT = -k\frac{dN}{N} \tag{16}$$

where:

k – constant,

After the integration from T_1 to T and from N_1 to N we will get:

$$T - T_1 = -k \ln N + k \ln N_1 \tag{17}$$

After arrangement and change to a decadic logarithm:

$$\log N = A - KT \tag{18}$$

where:

A and K are the constants

$$K = \frac{2{,}3}{k} \tag{19}$$

The relation (18) corresponds to actually measured results according to Figure 8, where the logarithm of cycles is N in the temperature of the die casting in the casting of single alloys. [14, 17]

4 Structural Analysis of Aluminium Castings

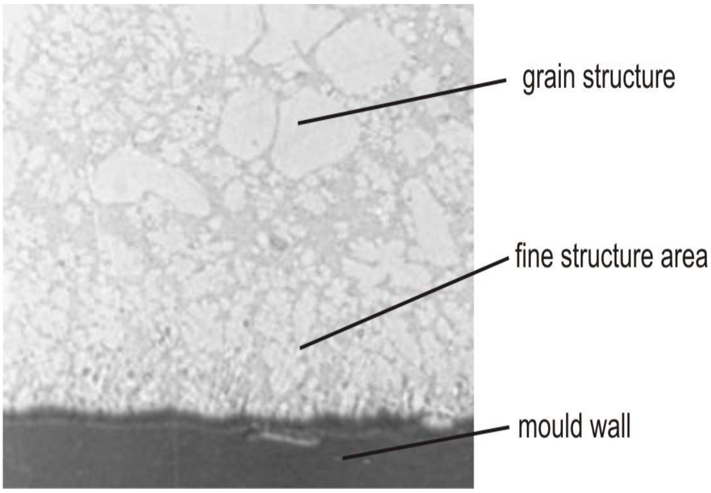


Figure 9
Microstructure of sample edge /250x/

For all analysed Al castings, the presence of extra fine crystals is typical, located in the area of the casting solidified at the face of a mould. The thickness of this area ranges from several µm to 1 mm. In the scratch pattern (Figure 9) the smooth transition from the fine crystals to the area with thicker structure can be observed. The formation of the area with fine structure relates to the degree of undercooling at the face of the mould, which depends mostly on the temperature of the mould. By contact between the melt flow and the cool wall of the mould, the temperature of which is lower than the one of the melt, crystallisation arises at a very high degree of undercooling and precisely limited direction of heat removal. Therefore, the area of fine crystals develops with primary orientation in the direction opposite to the heat removal direction. [8, 9]

The structure of aluminium castings of the analysed samples consists of α – solid solution (primary released) and eutectic (Figure 10). The eutectic appears to be a two-phase structure consisting of α – solid solution and silicon. Dark areas represent α – solid solution and the white ones represent the particles of eutectic silicon. Even with 2000-fold enlargement, the readability of structure aimed at the determination of silicon morphology is indistinctive. Therefore, the analysis of eutectic silicon morphology was made by means of a reflection electron microscope (REM), (Figure 11) with 6000-fold enlargement. With greater enlargement of the area, the presence of eutectic adhesion of individual eutectic cells can be seen, demonstrated in the picture by bright zigzag strips of various length. The eutectic is anomalous, i.e. the eutectic elements cannot grow by the same speed due to whatever causes [4, 5].

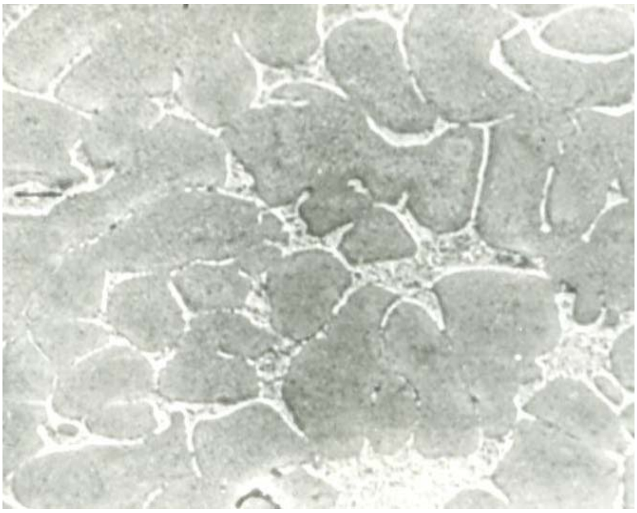


Figure 10
Basic structure (2000 x)

A fast growing phase is usually the phase with smaller volume rate. In the Al-Si system there are Si pseudo dendrites. The degree of undercooling necessary for the proceeding of the crystallisation queue for structural phases varies; therefore the width of present eutectic elements development area (i.e. area with a typical eutectic structure) depends on the gradient of the temperature of the border line between solid – liquid phase. Factors influencing the quality of a casting represented by mechanical characteristics of the casting are characteristics of α – phase and amount, size, and distribution of eutectic silicon. [6, 7]

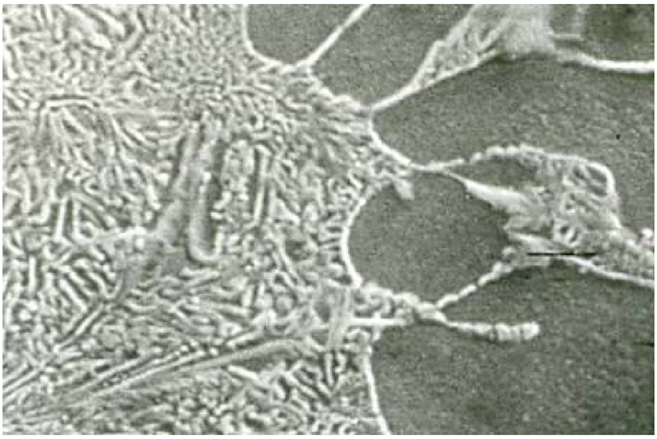


Figure 11
Adhesion of eutectic cells (6000x)

Conclusions

The presented work is aimed at the evaluation of the impact of thermal factors of die casting on the service life of the casting mould and the quality of pressurised Al-casting represented by the structural parameters of the castings, which the individual mechanical characteristics directly depend on [4, 5].

In die casting, after certain cycles, thermal fatigue strain of the mould arises. It leads to the formation of cracks and ends the service life of the mould according to the equation (7) that expresses the dependence of the number of mould service life cycles on the casting temperature in the casting of single alloys.

Further to the analysed scratch patterns of all samples it can be stated, provided that castings are produced from alloy of the same chemical structure, that the structural parameters are influenced mostly by the speed of cooling. The speed of cooling is primary controlled by the die casting temperature and the thermal content of the cast alloy (casting temperature). If the mould temperature is the same, the structural parameters are influenced only by temperature of the casting process [1, 3].

The structure and characteristics of the cast metals and alloys are therefore significantly influenced by the conditions of crystallisation. With technological interventions in the crystallisation process, the mechanical characteristics and structure may be improved. The crystallisation motion in practice is mostly influenced by the change in the speed of cooling. The increase of undercooling causes significant solidification of the melt under pressure due to the intensive heat exchange between the melt and the mould, which stimulates the development of areas with fine-grained crystals in the casting at the face of the mould, which smoothly moves to the area with thicker structure into the centre of the casting.

Acknowledgement

This article has been prepared within the project VEGA No. 1/0593/12: Research of Technological Parameters Influence of Die Castings and Design Modification of the Die System of the Casting Machine on Mechanical Properties of Die Castings of Lower Mass Category Made of Silumin.

References

[1] Belopuchov, A. K. a kol.: Casting of Metals under Pressure, Manufacturing of Engineering, Moskva, 1975

[2] Laudar, L.: Casting of Metals Under Pressure, SVTL Bratislava, 1964

[3] Malík, P., Gašpar, J., Paško, J.: Impact of Technological Factors of Pressure die Casting on Mechanical Properties of Castings) - 2009. In: Manufacturing Engineering. Roč. 8, č. 4 (2009), s. 32-37. - ISSN 1335-7972

[4] Ragan, E.: The Process of Casting Metals under Pressure, Technical University in Košice, FVT with seat in Prešov, Prešov 1997

[5] Valecký, J. a kol.: Casting of Metals under Pressure, STNL Praha, 1963

[6] Vinarcík, E. J.: High Integrity Die Casting Processes, John Wiley and sons, New York, 2003

[7] Gašpár, Š. – Maščeník, J. – Paško, J.: The Effect of Degassing Pressure Casting Molds on the Quality of Pressure Casting. In: Advanced Materials Research. Vol. 428 (2012), p. 43-46. - ISSN 1022-6680

[8] Bigoš, P. – Puškár, M.: Engine Output Increase of Two-Stroke Combustion with Exhaust System Optimization, In: Strojarstvo. Vol. 50, No. 2 (2008), pp. 69-76, ISSN 0562-1887

[9] Eperješi, Š. – Malík, J. – Vasková, I. – Eperješi, Ľ. – Fecko, D.: Comparison of Achieved Parameters Results of High-Strength Ductil Cast Iron by Different Way of Heat Treatment. In: Archives of Foundry Engineering. Vol. 11, special no. 1 (2011), pp. 55-57, ISSN 1897-3310

[10] Vajsová, V.: Optimization of Homogenizing Annealing for Al-Zn5.5-Mg2.5-Cu1.5 alloy. In. Metallurgist. Vol. 54, No. 9-10 (January 2011), pp. 618-622, ISSN 0026-0894

[11] Belov, N. A. - Belov, V. D. - Alabin, A. N. – Mishurov, S. S.: New Generation of Economically Alloyed Aluminum Alloys. In: Metallurgist. Vol. 54, No. 7-8 (November 2010), pp. 409-414, ISSN 0026-0894

[12] Zuberová, Z. – Sabirov, I. – Estrin, Y.: The Effect of Deformation Processing on Tensile Ductility of Magnesium Alloy AZ 31. In: Metallic Materials. Vol. 49, No. 1 (2011), pp. 29-36, ISSN 0023-432X

[13] Yin, D. L. - Weng, L. K. – Liu, J. Q. – Wang, J. T.: Investigation of Microstructure and Strength of AZ80 Magnesium Alloy by ECAP and Aging Treatment. In: Metallic Materials. Vol. 49, No. 1 (2011), pp. 37-42, ISSN 0023-432X

[14] Matisková, D.: Economic Reasons for Automatic of Component Production / - 2011. In: Manufacturing Engineering. N. 3 (2011), s. 56-58, ISSN 1335-7972

[15] Rózsa, Z.: Enterprise - a System with a Target Behavior In: Sedlák, M.: Business of Economy Bratislava: Iura Edition 2010, ISBN 978-808-8078-317-4

[16] Šebej, P., Hrubina, K., Wessely, E.: Creation of Production Planning Using the Mathematical Model and Multi-Criterion Optimal, In: Annals of DAAAM for 2004, Vienna: DAAAM International, pp. 413-414, ISBN 3901509429

[17] Modrák, V.: Functionalities and Integration Possibilities of Manufacturing Execution Systems, 2009. In: Annals of Faculty of Engineering Hunedoara - Journal of Engineering. Vol. 7, No. 1 (2009), pp. 51-56, ISSN 1584-2665, http://annals.fih.upt.ro/ANNALS-2009-1.html

Modeling and Analysing the Tensile Behavior of Fabric Samples

László Mihály Vas[1], Fatma Göktepe[2], Péter Tamás[3], Marianna Halász[4], Dicle Özdemir[5], Lívia Kokas Palicska[6], Norbert Szakály[7]

[1] Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Műegyetem rkp. 3, H-1111 Budapest, Hungary, vas@pt.bme.hu

[2] Namik Kemal University, Çorlu Engineering Faculty, Department of Textile Engineering, Silahtar Mahallesi, Sinan Dede Mevkii, Çerkezköy Yolu, 3. km, 59860 Çorlu, Tekirdağ, Turkey, fgoktepe@mmf.sdu.edu.tr

[3] Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Műegyetem rkp. 3, H-1111 Budapest, Hungary, tamas@inflab.bme.hu

[4] Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Műegyetem rkp. 3, H-1111 Budapest, Hungary, halaszm@pt.bme.hu

[5] Süleyman Demirel University, Engineering & Architecture Faculty, Department of Textile Engineering, 32260 Çünür, Isparta, Turkey, dicle@mmf.sdu.edu.tr

[6] Óbuda University, Rejtő Sándor Faculty of Light Industry and Environmental Protection Engineering, Doberdó ú. 6, H-1034 Budapest, Hungary, kokas.livia@rkk.uni-obuda.hu

[7] Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Műegyetem rkp. 3, H-1111 Budapest, Hungary, szakaly@mogi.bme.hu

Abstract: This paper presents a structural-mechanical model for describing the tensile behavior of textile fabrics in main directions based on the fiber-bundle-cells modeling theory and method. The applicability of this model, created by a variable transformed E-bundle shifted along the deformation axis, was demonstrated by analyzing the tensile load-deformation behavior and the breaking process including some size effects of a plain woven fabric made of OE rotor cotton yarns.

Keywords: tensile behavior of textile; modeling of textile; structural mechanical model; fiber-bundle-cells modeling; size effect of fabric samples

1 Introduction

Fibrous structures such as textile materials, fiber reinforced composites, and linear polymers are built up of discrete fiber-like elements such as textile or reinforcing fibers or yarns. The adjoining fibers or those intersecting a cross section of a fibrous sample create certain small assemblies that are fiber bundles in which the fibers show collective group-behavior [1-9]. The fiber bundle can be treated as intermediate elements of a fibrous structure, which can represent the statistical properties of the geometry or the strength.

In addition to the classic one [1], L.M. Vas et al. [4-9] have introduced some other idealized statistical fiber bundles called fiber-bundle-cells (FBC), and developed a modeling method as well as a software called FiberSpace [10-16], and shown that they can be applied to modeling some structural and strength properties of fibrous materials.

In this paper, a FBC model for describing the tensile behavior and breaking process of textile fabrics in main directions is presented and demonstrated in the case of a plain fabric made of OE rotor cotton yarns.

2 Fiber Bundle Cells-based Modeling Method

Statistical Fiber Bundle Cells

Fibers in a fibrous structure can be classified according to their geometry (shape, position) and mechanical behavior (strain state, gripping). These fiber classes are called fiber bundle cells (FBCs) (Fig. 1) [4-9].

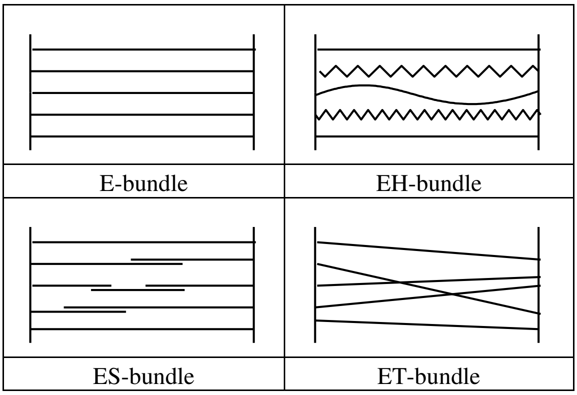


Figure 1

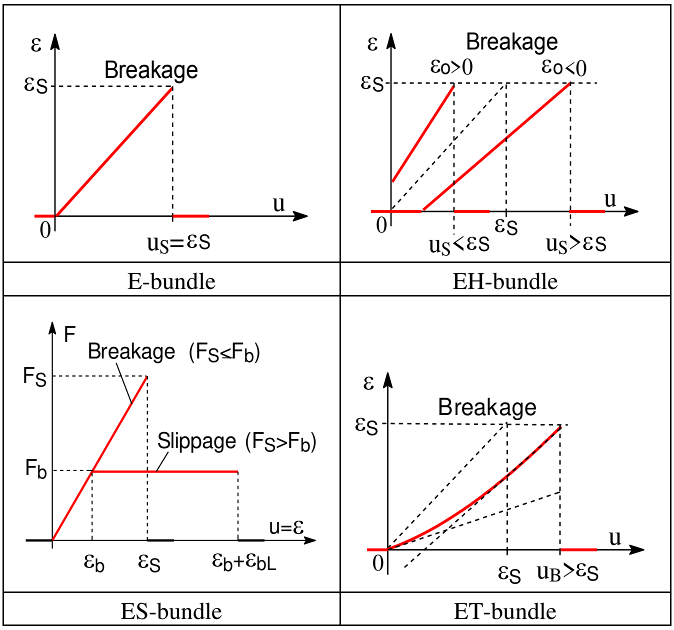
Structural scheme of the idealized fiber bundle cells

Fibers of these FBCs are supposed to be perfectly flexible and linearly elastic and to break at a random strain (ε_S). They are straight in the E-bundle, loose (ε_o<0) or pre-tensioned (ε_o>0) in the EH-bundle, and oblique (fiber angle $\beta \neq 0$) in the ET-bundle, and they are gripped ideally in these cases. Fibers in the ES-bundle are straight but they may slip out of their grip at a strain level ($\varepsilon_b < \varepsilon_S$) or create fiber-chains with slipping bonds. The shape, position, and strength parameters of fibers are assumed to be independent stochastic variables.

Considering a constant rate elongation tensile test, the tensile force ($F(u)$) creates a stochastic process as a function of the bundle strain (u). Being aware of the relationship between the bundle (u) and fiber strains (ε), the expected value of the tensile force of the FBCs ($E(F) = \overline{F}$) can be calculated as the sum of the single fiber forces using the suitable formulas developed [4-9]. Dividing the expected value by the mean breaking force of the fibers, the normalized tensile force of a bundle is computed as follows:

$$0 < FH(z) = F(z)/nF_S \le 1, \qquad z = u/\varepsilon_S \tag{1}$$

where n, $\overline{F}_S$, and ε_S are the number, the mean breaking force and strain of fibers, respectively, and z is the bundle strain normalized by the mean breaking strain of the fibers. Fig. 2 shows the graphic relationship between the strain of individual flexible fibers and the bundle.



E-bundle

EH-bundle

ES-bundle

ET-bundle

Figure 2

Relationship between the strains of single fibers and FBCs

In the case of an ES-bundle, ε_{bL} is the relative slippage way of fibers. In Fig. 3, the typical normalized expected value processes calculated at different parameter values are plotted for the FBCs. For the numerical calculations, all random parameters were assumed to be of normal distribution.

From Fig. 3 it is obvious that the FBCs can model rather complicated mechanical behaviors such as the initial convex part caused by crimped fibers (EH-bundle) or the slippages generated plateau beyond the peak (ES-bundle) even if they are used in themselves.

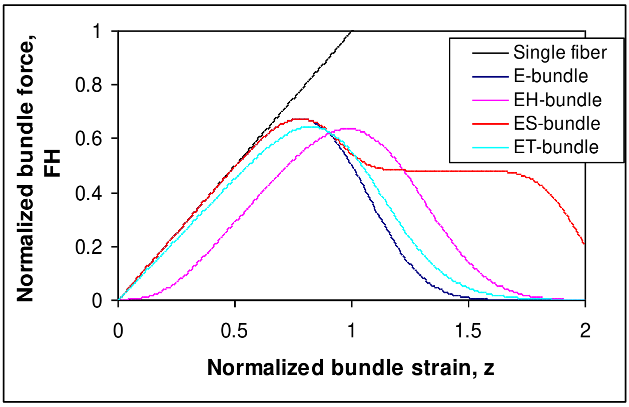


Figure 3

Expected value of typical normalized force-strain curves of the FBCs

Parallel and Serial Connection of Fiber-Bundle-Cells as FBC Models

In general, several types of FBCs are needed to model the response of a real fibrous structure. In most cases the parallel connected FBCs (Fig. 4a), called a composite bundle, provide a suitable model, and the resultant expected value process is calculated as the weighted sum of the single FBC responses where the weights are the fiber number ratios [4-10]. In the case where the size effect, such as the gauge length on strength, are examined, serial connection of the same type of independent FBCs is suitable to use in creating a bundle chain (Fig. 4b) [9, 11, 16].

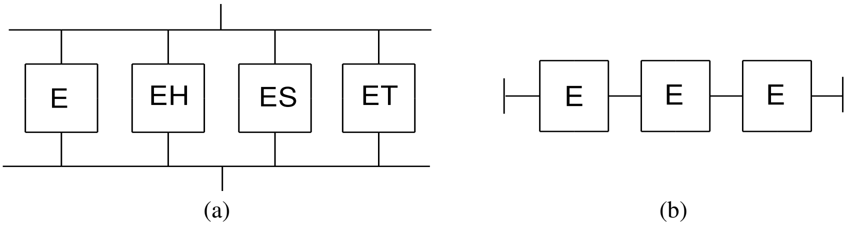
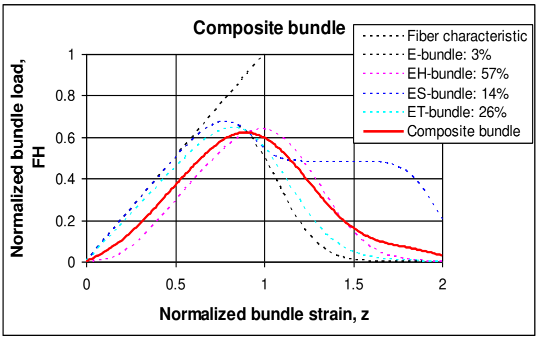


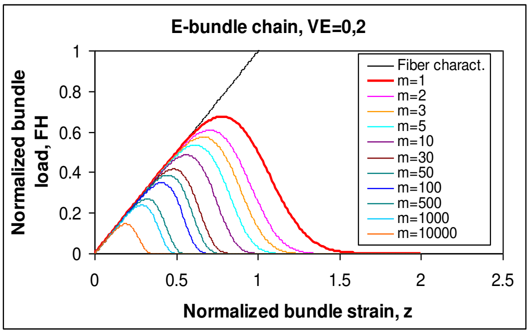
Figure 4

(a) Parallel and (b) serial connections of FBCs

As the examples in Fig. 5a show, the weighted sum of the normalized force-strain curves visible in Fig. 3 (percentages are the relative weight values of FBCs), while in Fig. 5b, the effect of the number (m) of serial connected E-bundles is demonstrated, causing the decrease of the peak value of the resultant force-strain curves that characterizes the tensile strength of the E-bundle chain (VE is the relative standard deviation of ε_S).



(a)



(b)

Figure 5

Normalized mean force-strain curves of parallel (a) and serial (b) connected FBCs

3 Structural Properties of Fabric Samples

Let us consider a rectangle sample with length L_o and breadth B_o cut out of the fabric in direction α where $\alpha=0^o$ and $\alpha=90^o$ are for the weft and warp directions, respectively (Fig. 6). Consequently, α is the angle of weft yarns to the length direction of the sample and the warp yarns are perpendicular to the weft direction ($\alpha+\beta=\pi/2$).

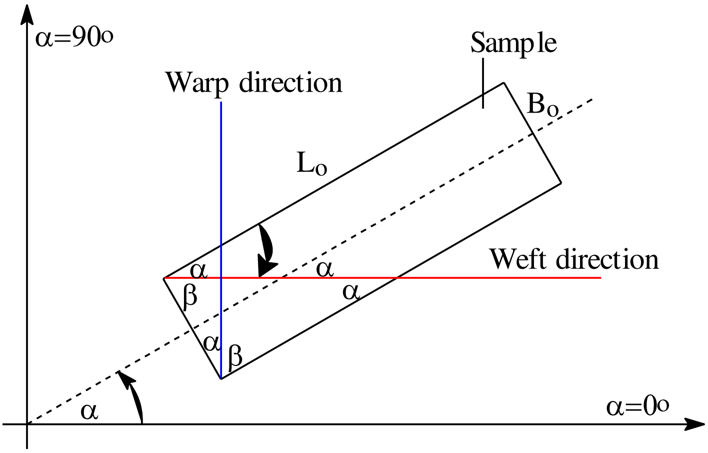


Figure 6
Disposition of a sample to the main structural directions of the fabric

The fabric sample is set for tensile test where it is gripped at the ends of B_o breadth, and L_o (assuming $B_o \le L_o$, as is usual in practice) is the free or gauge length of it; that is, L_o does not include the parts needed for gripping, as can be seen in Fig. 7.

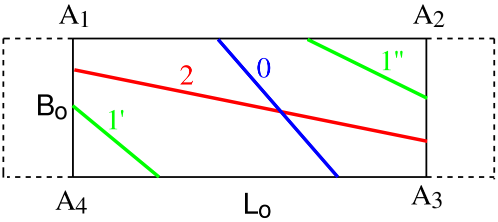


Figure 7
Classification of yarns according to their gripping position

The edges of the gripped part of the specimen bounding the free length, $\overline{A_1A_4}$ and $\overline{A_2A_3}$, can be called gripping lines. The yarns in the sample show different mechanical behavior according to the number of their gripped ends at the gripping lines (Fig. 7):

- 2-gripped yarns are gripped at both of their ends;
- 1-gripped yarns are gripped at one of their ends;
- 0-gripped yarns are not gripped at any of their ends.

4 Concept of Modeling Fabric Samples Using Fiber-Bundle-Cells

As a first step in the modeling, the tensile behavior of samples cut out in the main structural directions of the fabric will be tested and analyzed by using the FBC modeling method neglecting the crimping of yarns, as in the usual finite element layer models.

The expected value of the tensile force process of the E-bundle (Figs. 1-3) related to a single yarn (in the direction of the tensile load) can be calculated by the following formula [4]:

$$F(u) = E[F(u)] = \bar{K}u\left(1 - Q_{\varepsilon_S}(u)\right) \tag{2}$$

where u is the bundle strain, $\bar{K}$ is the mean tensile stiffness of the yarns, and $Q_{\varepsilon s}$ is the distribution functions of the breaking strain. The yarn parameters can be determined by tensile tests of yarn samples of gauge length L_o. Using the formulas according to Equation (1) for normalizing Equation (2), we obtain:

$$FH(z;L_o) = z\left(1 - Q_{\varepsilon_S}\left(z\varepsilon_S(L_o)\right)\right) = \bar{\kappa}(z)RH(z;L_o) \tag{3}$$

where the expected tensile characteristic ($\bar{\kappa}$) and the reliability function (*RH*) of the E-bundle are defined by Equation (3).

All this is valid for gauge length L_o at which the tensile test is performed, and it is well known that the tensile strength parameters of yarns depend on the gauge length [2, 3]. Supposing the gauge length is changed for $L=nL_o$ (n=1,2,…) and the section of fibers of length L_o create a so called bundle chain of independent elements (Fig. 4a), then the normalized expected tensile process of an E-bundle created by fibers of length L can be calculated as follows [16] (Fig. 5b):

$$FH(z;L) = z\left(1 - Q_{\varepsilon_S(L)}\left(z\varepsilon_S(L)\right)\right) = z\left(1 - Q_{\varepsilon_S(L_o)}\left(z\varepsilon_S(L_o)\right)\right)^{L/L_o} \tag{4}$$

It can be noted that formula (4) can be extended for $L<L_o$ as well. Consequently, the following relationship can be obtained, which is valid for both $0<L<L_o$ and $L\geq L_o$:

$$\left(\frac{\mathrm{FH}(z;\mathrm{L_o}/\mathrm{n})}{\mathrm{z}}\right)^{\mathrm{n}} = \frac{\mathrm{FH}(z;\mathrm{L_o})}{\mathrm{z}} \tag{5}$$

According to both modeling and experience, the mean breaking force of yarns and its standard deviation increase with a reduction in the gauge length, which is known as size effect in the literature [3, 16].

5 Application of the FBC Modell

5.1 Experimental

Material Tested

To demonstrate the applicability of the FBC model of fabric samples, a plain woven cotton fabric and its yarn components were tested (Table 1).

Table 1

Data of the examined fabric

Material	Type		Weight [g/m^2]	Type of weave	Yarn count [tex]		Number of yarns λ [1/cm]		Twist direction of yarns	
	warp	weft			warp	weft	warp	weft	warp	weft
Cotton	OE rotor	OE rotor	156	Plain	29.6	29.6	26	22	Z	Z

Tensile Tests Results

The weft and warp yarns were examined by tensile testing using a gauge length of 50 mm, as is used normal, e.g. on the KES System. The test speed and the pretension were 12 mm/min and 0.2 cN, respectively. The results are summarized in Table 2.

Table 2

Tensile test result of yarn

Statistical properties	**Yarn** (29.6 tex, Z twist)		
	Co-ordinates of the breaking point		
	Force [cN]	**Elongation** [mm]	**Strain** [%]
Mean	272.4	3.24	6.5
S.D.	32.3	0.35	0.7
C. of V. [%]	11.9	10.76	10.8

Finally, tensile tests were carried out on fabric samples (Table 3) cut out in main directions (weft and warp) with a breadth of 50 using tensile tester Zwick Z50. The gauge length and the test speed were 50 mm and 12 mm/min, respectively.

The strength data are summarized in Table 3, where MaxF denotes the peak value of any load-elongation curve that is the breaking force, and the averaged curve was calculated by averaging the single load-elongation measurements point by point.

Table 3
Tensile test results of fabric samples

Fabric P - Breadth: 50 mm				
	Measured			Averaged curve
	weft_1	weft_2	weft_3	
MaxF [N]	418,5	432,8	412,9	**380,9**
Δl(maxF) [mm]	7,31	8,05	7,82	7,31
ε(maxF) [%]	14,62	16,10	15,64	14,62
				Averaged curve
	warp_1	warp_2	warp_3	
MaxF [N]	492,0	480,0	500,7	**472,4**
Δl(maxF) [mm]	8,36	8,83	8,82	8,41
ε(maxF) [%]	16,72	17,66	17,63	16,82

5.2 Results of FBC Modeling

The results of tensile measurements performed on fabric P in the main structural directions and its yarn component form the experimental background of the FBC modeling.

Cutting out a sample in weft direction from the fabric means that the sample is built of weft yarns aligned lengthwise and warp yarns aligned crosswise (Fig. 8b). Loading this sample in lengthwise direction, the load is taken up by the 2-gripped weft yarns, and the 0-gripped warp yarns play just a modifying role by interlacing the weft yarns even if densely. In the present paper, the effect of interlacing is taken into account as a kind of adhesion between the yarns, that plays an important role in the cases of 1- or 0-gripped yarns, but it can be neglected for the 2-gripped yarns.

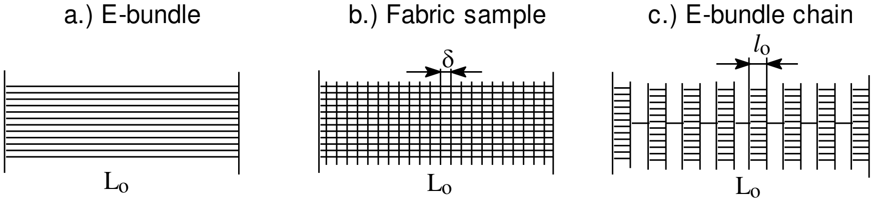


Figure 8
E-bundle (a), a fabric sample cut out in main direction (b) and E-bundle chain (c) as the model of the sample

Consequently, in a rough view, the sample cut out in main direction creates a bundle of 2-gripped yarns with an orientation angle of zero; therefore it can be modeled by a simple E-bundle (Fig. 8a).

The normalized expected tensile force process related to a single yarn of an E-bundle is given by Equation (3). The number of yarns loaded is determined by the yarn densities (λ_{ok}, k=1,2) (Table 2), which by multiplying with formula (3) provides the FBC estimation of tensile force recorded during a tensile test of a sample cut out in the weft (k=1) or warp (k=2) direction:

$$\begin{aligned} E[F(u)] &= E[F_k(u)] = \lambda_{ok} B_o E[F^{(k)}1(u)] = \\ &= \lambda_{ok} B_o \bar{F}_S^{(k)} z\left(1 - Q_{\varepsilon_S}^{(k)}\left(\varepsilon_S^{(k)} z\right)\right) = \lambda_{ok} B_o \bar{F}_S^{(k)} FH^{(k)}(z) \end{aligned} \tag{6}$$

This is valid in the case of an E-bundle built up of yarns of given length L_o.

Modeling Software FiberSpace is suitable for providing $FH^{(k)}(z)$ because this simple model does not contain any combined FBC. For modeling – in this simple case – just the mean and CV of the breaking strain of the yarn are needed, which can be found in Table 2. They are respectively denoted by AE (=0.0648) and VE (=0.108) in Fiber Space.

The results of modeling the E-bundle of L_o=50 mm yarns – which was imported to Microsoft Excel – can be seen in Fig. 10 ($L=L_o$=50 mm). This expected tensile force process can be approximated by averaging point-by-point measured and normalized force-strain curves. According to Fig. 10, the expected yarn strength efficiency (which is determined by the peak value of the normalized curve) in the fabric sample in main direction is 0.782, that is 78.2%. This is valid for both weft and warp directions because of the identical weft and warp yarns and the symmetric structure of a plain weave.

The estimated expected tensile strength in both directions can be calculated by using Equation (6). As is usual according to the related standards, the breadth of the sample was taken as B_o=50 mm (Table 4).

The strength values in Table 4 are considered 'ideal' when they were calculated as the simple product of the number of yarn in the direction of load and the mean yarn strength, and the 'realistic' ones were obtained by multiplying the latter by the expected yarn strength utilization provided by modeling the bundle of yarns of 50 mm length (Fig. 8a).

On the basis of Table 4, it can be stated that the measured tensile strength values proved to be larger than the estimations denoted by 'ideal' or 'realistic'. This can be explained by two facts:

(1) The mean tensile strength of yarns strongly depends on the gauge length used (size effect); the smaller it is, the larger the mean strength is [1-3, 16].

(2) The crosswise yarns create a kind of gripping for the tensile loaded yarns, sectioning them into short E-bundles which form a so-called E-bundle chain (Fig. 8c) [16]. The effective length of these bundles (l_o; Fig. 8c) can be larger than the distance between the crosswise yarns (δ; Fig. 8b) because of the possibility of some slippage and the strain of yarns at the peak force.

Table 4

Results of FBC modeling using yarns of 50 mm length and measurements (mean)

	Estimation	Weft direction	Warp direction
Yarn strength utilization [%]	real	78.2	78.2
Fabric tensile strength calculated for 50 mm width [N]	ideal	300	354
	realistic	234	277
Fabric tensile strength measured at 50 mm width [N]	----	381	472

The mean strength of these short E-bundles can be much greater than that of the longer, but the standard deviation of these short yarn segments is larger as well. The strength of this bundle chain is determined by the "weakest link" [2], yet this minimum value can be significantly larger than that of the original bundle of long yarns.

Regarding the breaking force only, suppose all the yarn breakages take place in a single short bundle (Fig. 9a), meaning that the other bundles are subjected to strain only and the model of this behavior can be represented by a short E-bundle and a serial connected elastic continuum part (Fig. 9b).

In this case, the force-elongation relation is governed by the E-bundle, and the role of the elastic part is to model the surplus in elongation as the contribution of the other bundles. In this model, reaching the peak force value of the bundle, the breakage of the chain can occur in a catastrophic way if the breaking bundle cannot cover the loss in elongation after the bundle force peak [16]. These drops in force can take place after each other if the yarn breakages are distributed over several bundles of the chain.

As the other extreme case, the breakages of single yarns can be evenly distributed over the chain, realizing an expected tensile process identical with that of the single bundles [16].

In reality, the damage and failure process is realized as one between the two extreme damage cases. In addition, the yarn chain that is a bundle chain which consists of a single yarn can be treated as a lower estimation of the real one [16].

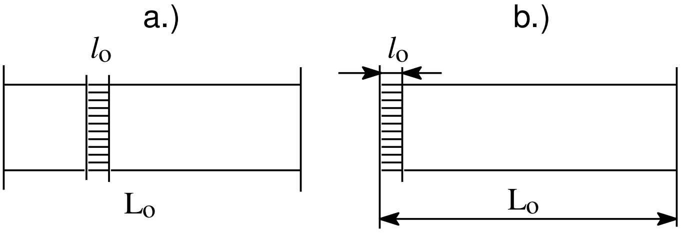


Figure 9

Fabric sample as the serial connection of a single breaking E-bundle and an elastic part

In the sense of the “weakest link” concept, Equation (4) describes the expected tensile force process of an E-bundle chain where the number of E-bundle is $n=L/L_o$. In this case, the peak value and the half-width of the force-chain curve (corresponding to the standard deviation of the yarns) decrease by increasing n or L (Fig. 10). At the same time, concerning the tensile strength, the E-bundle chain can be considered as a single E-bundle built up of yarns of length L. In this sense Equation (4) can be used for bundles of yarns of length smaller than L_o as well. In the latter case, the peak value and the half-width increase (Fig. 10).

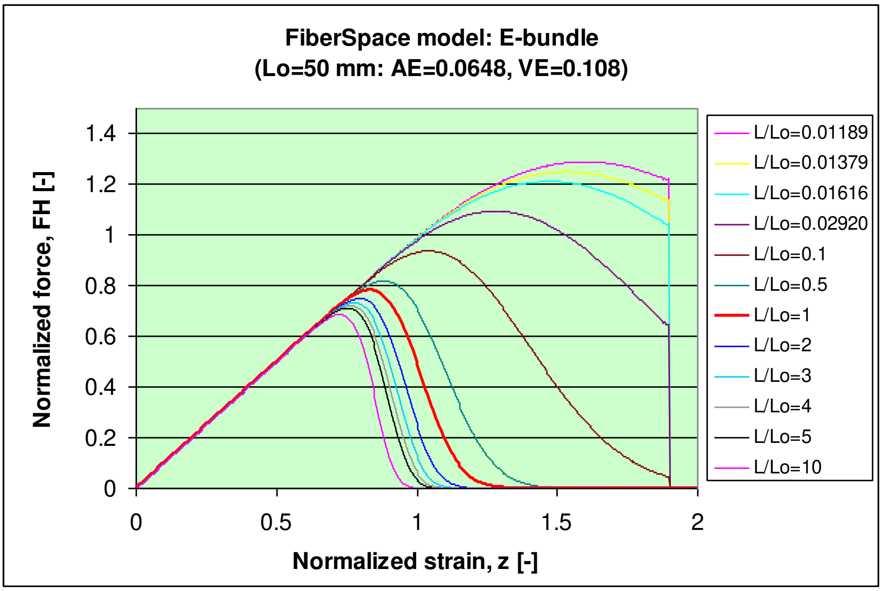


Figure 10

Normalized expected force-strain curves of E-bundles with different relative lengths (L/L_o)

In this extended sense, Fig. 11 shows the normalized peak force values as a function of the yarn length, which is the gauge length of the yarns (L) in logarithmic scale.

The results of modeling discussed above give a good basis for analyzing the measurements obtained by tensile test of fabric samples. Since our simple FBC modeling method applies E-bundles only consequently, the so called structural elongation caused by the crimping of the yarns and the elastic pulling out of the grips are not modeled, and therefore the first step of the analysis is the determination of the structural elongation.

This is defined by the steepest tangent straight line belonging to the inflexion point of the rising part of the force-elongation curve. The structural elongation is determined by the intersection point of the tangent and the elongation axis (Fig. 11).

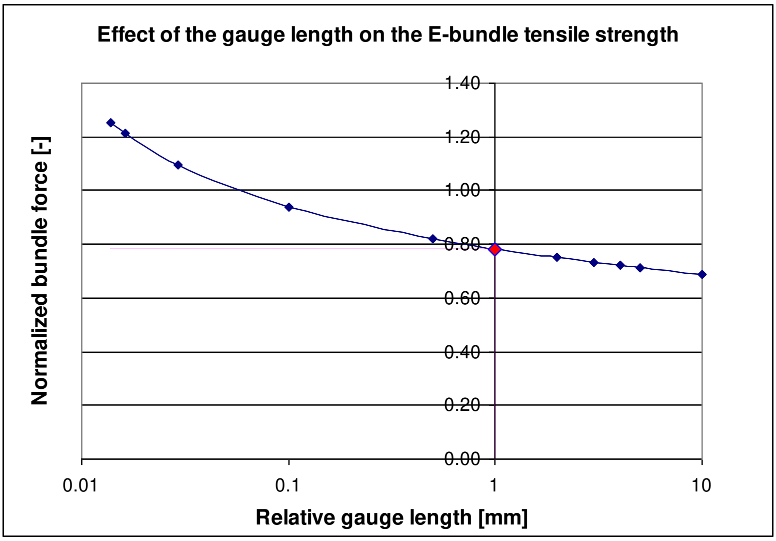


Figure 11

Peak values of E-bundles versus relative yarn length in logarithmic scale

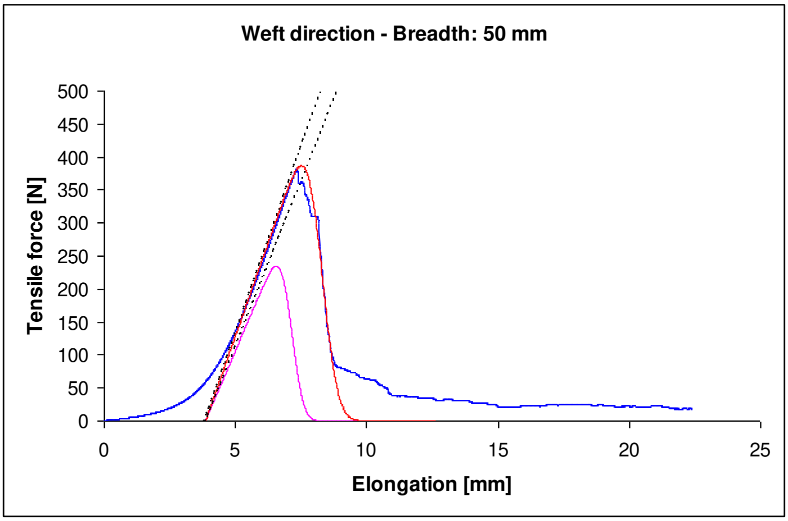


Figure 12

Measured and averaged force-elongation curve (blue line), shifted E-bundle curve (lilac line), and the transformed model curve (red line) for samples cut out of weft direction

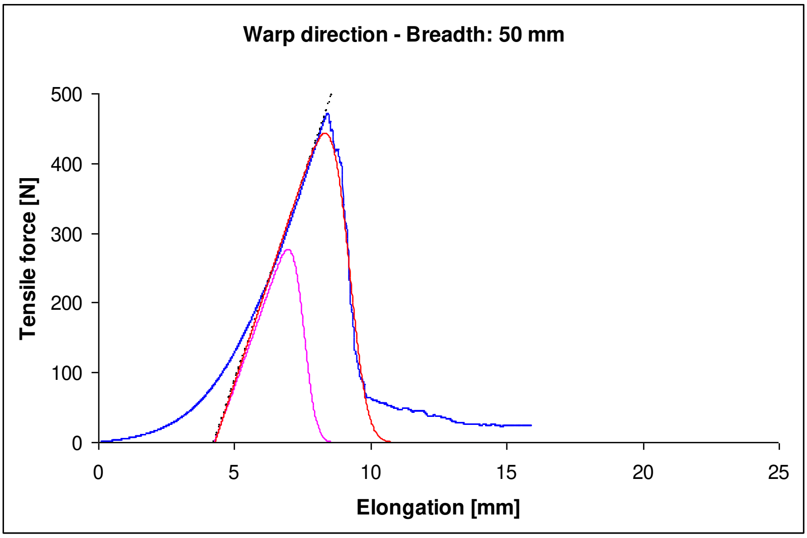


Figure 13

Measured and averaged force-elongation curve (blue line), shifted E-bundle curve (lilac line), and the transformed model curve (red line) for samples cut out of warp direction

Table 5

Results of FBC analysis based on E-bundle model

Sample width L_o [mm]	**Origin of result**	**Properties**	**Weft**	**Warp**
50	**Measured**	Size of weave cell δ [mm]	0.4545	0.3846
		Tensile strength [N]	380.9	472.4
		Tensile stiffness [N/mm]	113	115
		Structural elongation [mm]	3.85	4.26
	Modeled by yarns	Max force [N]	234.4	277.0
		Tensile stiffness [N/mm]	100	115
		Yarn strength utilization [%]	78.21	78.21
	Shifted and scaled model	Scale factor of elongation [-]	1.35	1.50
		Scale factor of force [-]	1.65	1.60
		Max force [N]	386.7	443.2
		Yarn strength utilization [%]	1.29	1.2514
		Length ratio, $n=l_o/L_o$	0.01190	0.01379
		Effective bundle length, l_o [mm]	0.5945	0.6897
		Relative eff. bundle length, l_o/δ [-]	1.31	1.79

Figs. 12 and 13 show the measured and averaged force-elongation curves (blue lines) and the steepest tangents.

In these diagrams the E-bundle model curves (lilac lines) with their initial tangent are shifted from the origin by the structural elongation.

It can be seen in the diagrams (Figs. 12, 13) that the linear variable transformation (which is applying the proper scaling) of these shifted E-bundle curves (red lines) fits well to the measured ones regarding both the rising and the falling branches of the curves.

The results of the measurements and modeling and model based analysis are summarized in Table 5.

Conclusions

On the basis of the diagrams and numerical results some essential statements can be made.

(1) The structural elongation caused by the interlacing and crimping of the yarns adding to that the elastic pulling out of the grips is rather large (about 8%).

(2) The large measured structural elongation means that it is important to take into account for modeling the deformation and, e.g., the drape behavior of the fabric.

(3) The yarn strength utilization of ideal E-bundle consisting of independent yarns of 50 mm length is 78.2%, which is relatively small. The measured utilization was larger than 100%, meaning that the interlaced crosswise yarns bind together the segments of the loaded yarns, forcing them strongly to work together, and by that, a relatively small effective length is realized that is much smaller than the gauge length of the fabric sample.

(4) The effective bundle length values (l_o) determined by the shifted and rescaled E-bundle curves using Equation (4) are larger than the weave cell sizes (δ), indicating that the crosswise interlacing yarns can slip on the loaded ones causing an increase in the bundle length.

(5) The slippage of interlacing yarns given by the positive difference l_o-δ (see Table 5), which is in relation with certain friction and shear effects, indicates that it can also be an important factor in modeling the deformation and drape behavior of fabrics.

On the basis of these results, the E-bundle based modeling of the tensile strip test of fabrics in the main directions can be well used for analyzing the tensile measurements and the structural-mechanical behavior of fabric specimens tested.

Acknowledgements

This work was supported by the Hungarian Scientific Research Fund (OTKA, H), the National Development Agency (NDA, H), and the Scientific and

Technological Research Council of Turkey (TÜBİTAK, TR) since this multinational study has been carried out commonly as part of the project K68438 (OTKA-NDA), as well as of the S&T projects 108M604 (TÜBİTAK) and TR-17/2008 (NDA) respectively. This project is also supported by the New Széchenyi Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002) and connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project.

References

[1] Harlow D. G., Phoenix S. L.: *The Chain-of-Bundles Probability Model For the Strength of Fibrous Materials I: Analysis and Conjectures*. Journal of Composite Materials Vol. 12, 195-214, and *The Chain-of-Bundles Probability Model For the Strength of Fibrous Materials II: A Numerical Study of Convergence*. Journal of Composite Materials Vol. 12, 314-334 (1978)

[2] Peirce F. T.: *Tensile Tests for Cotton Yarns. Part V The Weakest Link, Theorem on the Strength of Long and Composite Specimens*. Journal of The Textile Institute Transactions 17(7) T355-T368 (1926)

[3] Sutherland L. S., Shenoi R. A., Lewis S. M.: *Size and Scale Effects in Composites: I. Literature Review*. Composites Science and Technology 59, 209-220 (1999)

[4] Vas L. M., Császi F.: *Use of Composite-Bundle Theory to Predict Tensile Properties of Yarns*. Journal of the Textile Institute 84(3), 448-463 (1993)

[5] Vas L. M., Halász G.: *Modelling the Breaking Process of Twisted Fibre Bundles and Yarns.* Periodica Polytechnica 38(4), 297-324 (1994)

[6] Vas L. M.: *Strength of Unidirectional Short Fiber Structures as a Function of Fiber Length*. Journal of Composite Materials 40(19), 1695-1734 (2006)

[7] Vas L. M., Rácz Zs.: *Modeling and Testing the Fracture Process of Impregnated Carbon Fiber Roving Specimens During Bending Part I. Fiber Bundle Model*. Journal of Composite Materials 38 (20), 1757-1785 (2004)

[8] Vas L. M.: *Statistical Modeling of Unidirectional Fiber Structures*. Macromolecular Symposia. Special Issue: Advanced Polymer Composites and Technologies 239(1), 159-175 (2006)

[9] Vas L. M., Tamás P.: *Modelling Method Based on Idealised Fibre Bundles*. Plastics, Rubber and Composites 37(5/6) 233-239 (2008)

[10] Molnár K., Vas L. M., Czigany T.: *Determination of Tensile Strength of Electrospun Single Nanofibers through Modeling Tensile Behavior of the Nanofibrous Mat*. Composites Part B 43 (2012) 15-21 DOI: 10.1016/j.compositesb.2011.04.024 (2012)

[11] Vas L. M., Tamás P.: *Fiber-Bundle-Cells Method and its Application to Modeling Fibrous Structures*, GÉPÉSZET 2006, 5th Conf. on Mech. Eng. Budapest, May 25-26, 2006. Proceedings (CD – Full-text) ISBN 963 593 465 3 (2006)

[12] Vas L. M., Tamás P.: *Modeling Fibrous Reinforcements and Composites Using Fiber Bundle Cells*, III. International Technical Textiles Congress Dec. 1-2, 2007 Istanbul, Proceedings 95-104 ISBN 978-975-441-245-1 (2007)

[13] Tamás P., Vas L. M.: *Modelling Fibrous Structures by FiberSpace Using Fourier Transformation Based Expert Program*. GÉPÉSZET 2008, 6th Conf. on Mech. Eng. Budapest, May 28-29, 2008, Proceedings (CD – Full-text) ISBN 978 963 420 947 8 (2008)

[14] Tamás P., Vas L. M.: *Fiber Bundle Based Modeling Software*, ITC&DC Magic World of Textiles IV International Textile, Clothing & Design Conference Oct. 5-8, 2008 Dubrovnik, Book of Proceedings 892-897, ISBN 978-953-7105-26-6 (2008)

[15] Vas L. M., Tamás P.: *Modelling Size Effects of Fibrous Materials Using Fibre-Bundle-Cells*. ECCM-14 14th European Conference on Composite Materials, Budapest, 7-10 June, 2010, Proceedings Paper ID-705, 1-11, ISBN: 978-963-313-008-7 (2010)

[16] Vas L. M., Tamás P., Halász M., Göktepe F.: *Fiber-Bundle-Cells Model of Composites*. 5th Aachen-Dresden International Textile Conference, Aachen, Germany, Nov. 24-25, 2011 (poster presentation) Proceedings ISSN 1867-6405 (CD edited by B. Küppers) P-15, 1-10 (2011)

Compensation of the Impact of Disturbing Factors on Gas Sensor Characteristics

Zvezditza Nenova, Georgi Dimchev

Technical University of Gabrovo
Department of Electrical Engineering
4, H. Dimitar Str., Gabrovo 5300, Bulgaria
nenova@tugab.bg; gdimchev@tugab.bg

Abstract: Methods for gas control have been extensively developed for the monitoring of air quality, for gas leak control, for the development of 'electronic nose' systems, etc. Metal oxide gas sensors have been widely used in particular. However, apart from changes in the controlled gas concentration, changes in their parameters also depend on ambient conditions. The main impact comes from temperature and humidity. Therefore, the compensation of these disturbances is important for increasing the accuracy of concentration measurements of the controlled gases and the reliability of control. The present paper proposes a method for compensating the impact of temperature and humidity on gas sensor characteristics using artificial neural networks. This compensation method is applied to the control of methane concentration by gas sensors TGS813 and TGS2611. The results obtained confirm the applicability of this method.

Keywords: compensation; gas sensors; disturbing factors; artificial neural networks

1 Introduction

Gas systems are widely used for monitoring outdoor and indoor air quality, in gas leak control systems, in the chemical industry, in the development and implementation of 'electronic nose' systems, etc. [1-5]. The control of air parameters is important for the protection of the environment and human health, as well as for providing safe working conditions. Gas pollution can spread over a wide area in a short time, and therefore methods and equipment for its measurement and monitoring are being extensively developed. A wide range of gas sensors [6-10] have been designed, including metal oxide gas sensors. Different kinds of metal oxides such as SnO_2, ZnO, Fe_2O_3, WO_3, Co_3O_4, etc. [11-16] are used as sensing materials. Their operating principle is based on increasing the conductivity of the surface film of the sensitive element when the test gas is adsorbed. Depending on the composition of the surface film, the sensor responds to different gases such as carbon monoxide, carbon dioxide, ethanol, methane, propane, ammonia, hydrogen sulfide, hydrogen, etc. [6-11]. Metal oxide gas sensors have high sensitivity, low cost and a short response time. However, their characteristics are influenced by various ambient parameters which act as disturbing factors in gas control. Temperature and humidity have a major impact among these factors [7-10, 17, 18]. To enhance the measurement accuracy and reliability of control, compensating the impact of disturbing factors on gas sensors is of prime importance.

A method for compensating the impact of ambient temperature and humidity on gas sensor characteristics by using artificial neuron networks (ANN) is proposed in this paper. The method is based on a three-dimensional approximation of the gas sensor characteristics employing an ANN. The method is applied to the control of methane concentration with gas sensors TGS813 and TGS2611, and the results of that implementation are shown.

2 ANN Compensation Method

For metal oxide gas sensors, the input quantity is the unknown concentration, *Conc,* of the gas being controlled, which leads to a change in the output quantity of the sensor - its resistance, *Rs*. The ambient factors, temperature, *t*, and relative humidity, *RH*, which act as disturbing factors, also have an effect on this resistance (Fig. 1).

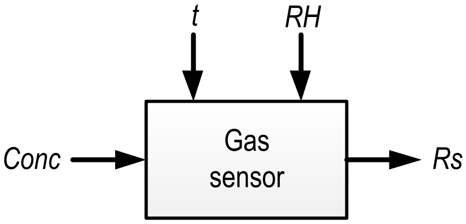


Figure 1
Input quantity *Conc* and disturbing factors *t* and *RH* for gas sensors

Sensor manufacturers usually report gas sensor characteristics as the sensor resistance ratio, Rs/Ro, under various gas concentrations and ambient conditions, i.e.:

$$\left(\frac{Rs}{Ro}\right)=f\left(Conc,\, t, RH\right) \quad , \tag{1}$$

where *Rs* is sensor resistance, and *Ro* is resistance for referent concentration, temperature and humidity.

However, these characteristics are usually given only for some values of the disturbing factors

$$\left(\frac{Rs}{Ro}\right)_i = f\left(Conc\right)\Big|\ t_i, RH_i = const \qquad (2)$$

$$i = 1,2, \ldots, n\ .$$

It should be noted that in practice it is difficult to calibrate the gas sensor, and the impact of disturbing factors is usually given only for characteristics at a fixed concentration.

$$\left(\frac{Rs}{Ro}\right)_i = f\left(t_i, RH_i\right)\Big|\ Conc = const \qquad (3)$$

$$i = 1,2, \ldots, n\ .$$

If in the application of gas sensors the operating characteristic is chosen for fixed *t* and *RH* (most commonly at 20°C/65%RH), this inevitably leads to measurement errors due to changes in ambience. In order to take into consideration the impact of *t* and *RH*, it is necessary to approximate the transformation function of the sensor and use it in applications.

Based on equation (1), this should be a three-dimensional approximation. Difficulties arise owing to the great nonlinearity of characteristics (1) – (3). Additionally, as was mentioned, the sensor characteristics usually cannot be given at uniform points that can be used in function approximation.

A theoretical method for polynomial approximation of a multivariable sensor characteristic was proposed in [18]. In its practical application only the compensation of the impact of humidity, *RH*, on gas sensor characteristics is shown. However, introducing a second disturbing factor (such as temperature), would substantially increase the number of equations and coefficients used.

Artificial neural networks can also be employed for solving different problems with many input parameters. It is very common to use ANN for gas and odor recognition, the classification of products, the control of environmental parameters, etc. [19-23]. In [24] an ANN-based virtual compensator for correcting the effect of a disturbing variable in transducers is proposed. That method is applied to a strain-gauge transducer-based pressure measurement system. The correction is carried out by a nonlinear two-dimensional artificial neural network-based inverse model of the transducer. ANN has also been used for two-dimensional approximation of humidity sensor characteristics in order to compensate for the impact of one factor - temperature [25]. This approach is shown to achieve the highest accuracy for a nonlinear transformation function compared to polynomial and interpolation methods. The compensation for temperature effects in gas sensors via ANN is reported in [26].

The ANN-based method proposed in this paper aims to compensate for the impact on gas sensors of both ambient temperature and humidity through a three-dimensional approximation of the gas sensor characteristics.

The method is implemented in two stages: training of the ANN, and real measurement and control of gas concentration.

In the training stage, the calibration characteristics given by the manufacturers are used. Input parameters for the ANN are: the gas sensor resistance ratio, $Rs\,/\,Ro$, ambient temperature, *t*, and relative humidity, *RH*, and an output parameter – the concentration, *Conc,* of the respective gas. The points of training can be complemented, whenever possible, by a functional approximation of characteristics (2) and offsets based on (3).

As a result of the ANN training, a three-dimensional approximation of the sensor characteristics is performed with relationships of the type

$$Conc = f\left(Rs\,/\,Ro,\, t,\, RH,\, W,\, a,\, b\right)\ , \tag{4}$$

where *W*, *a* and *b* are ANN parameters.

In the stage of real measurement and control, in addition to measuring the gas sensor parameter, it is necessary to measure temperature separately by means of a temperature sensor and air humidity by means of a humidity sensor. A schematic diagram of the method implementation for one gas is shown in Fig. 2.

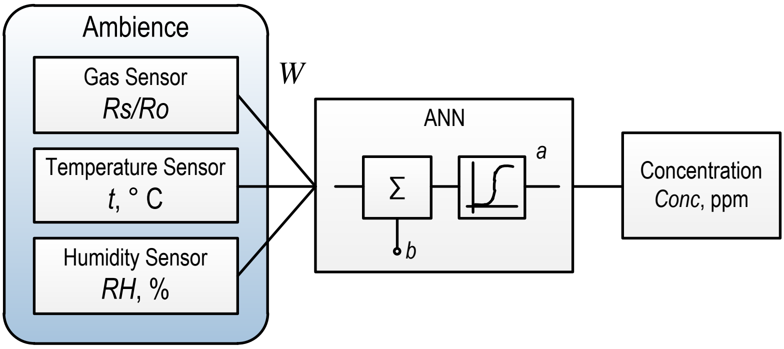


Figure 2

Schematic diagram of the implementation of the ANN compensation method for one gas

On the basis of the approximation relationships obtained (equation (4)) the measured gas concentration is determined. The changes in ambient temperature and humidity are taken into account, and therefore, their impact on gas sensor characteristics is compensated for.

The method can also be employed for a higher order approximation for a greater number of disturbing factors.

3 Results and Discussion

The method is applied to compensate for the impact of temperature and humidity on gas sensors TGS813 and TGS2611 for the control of methane concentration. Sensor characteristics (2) and (3) at 1000 ppm and 5000 ppm, respectively, given by manufacturers have been used [7].

For sensor TGS813, *Ro* is the gas sensor resistance for the referent concentration of 1000 ppm and 20°C/65%RH. According to experimental characteristics [7], in logarithmic scale, the characteristics $Rs/Ro = f(Conc)$ of the sensor for given t and RH are straight lines and can be represented by an equation of the form

$$y = a_0 + a_1.x \ , \tag{5}$$

where $y = \lg(Rs/Ro)$, $x = \lg(Conc)$.

These characteristics are parallel straight lines; i.e., coefficient a_1 is constant and can be determined by any of the experimental relationship $Rs/Ro = f(Conc)$ for $t = const$ and $RH = const$.

Variations in temperature and relative humidity lead only to a change in the offset a_0 of these characteristics. This offset has been calculated on the basis of characteristics (3) at the reference concentration for temperature variation in the range of -10°C …+40°C and relative humidity in the range of 0…100%RH [7].

Thus, the family of characteristics are obtained analytically at various temperatures in the range of -10°C …+40°C and fixed humidities of 0, 20, 40, 65 and 100%RH. Fig. 3 presents this family of characteristics for sensor TGS813 at 65%RH, showing the impact of temperature.

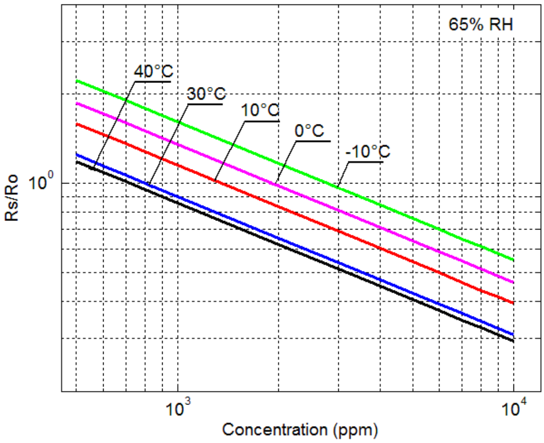


Figure 3

Analytically obtained characteristics for sensor TGS813 for temperature variation and 65%RH

Similarly, based on the experimental characteristics [7] for sensor TGS2611, the families of characteristics have been obtained at temperatures in the range of -10°C to 40°C and for fixed values of relative humidity of 35, 50, 65 and 95%RH. For this sensor, Ro is the resistance at a referent concentration of 5000 ppm and 20°C/65%RH. Fig. 4 presents the family of analytically obtained characteristics for sensor TGS2611 for temperature variation and 65%RH.

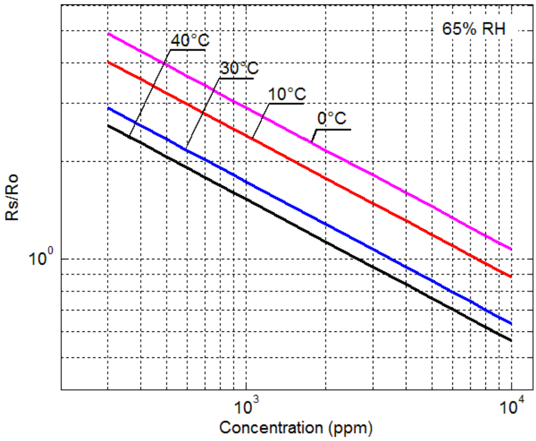


Figure 4

Analytically obtained characteristics for sensor TGS2611 for temperature variation and 65%RH

The experimental characteristics, apart from those at 20°C/65%RH, and the whole set of analytically obtained characteristics for each sensor TGS813 and TGS2611 is used for ANN training. The experimental characteristics at 20°C/65%RH are used for checking the accuracy of the proposed method.

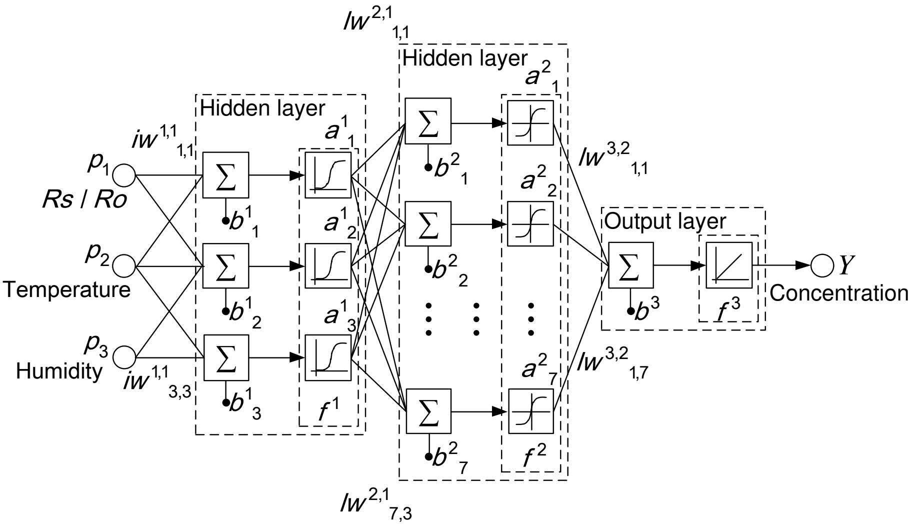
Experiments with various algorithms have been carried out for the ANN training. The best convergence for the smallest number of neurons is obtained in training with the LM algorithm (Levenberg-Marquardt back propagation). The obtained ANN with back propagation of error has three layers: two hidden (input, intermediate) and one output layer. The first layer consists of three neurons, one for each input quantity, the second layer is made up of seven neurons, and the third layer has one neuron (Fig. 5).

In both the first and second layers the transfer functions of neurons $\left(f^1\right)$ and $\left(f^2\right)$ are sigmoidal, and in the third layer $\left(f^3\right)$ it is linear.

The neural network has the following form

$$Y = f^3\left(LW^{3,2} f^2\left(LW^{2,1} f\left(IW^{1,1} p + b^1\right) + b^2\right) + b^3\right), \tag{6}$$

where $Y = Conc$, $p_1 = Rs / Ro$, $p_2 = t$, $p_3 = RH$.



$$a^1 = f^1(IW^{1,1}p + b^1) \quad a^2 = f^2(LW^{2,1}a^1 + b^2) \quad Y = f^3(LW^{3,2}a^2 + b^3)$$

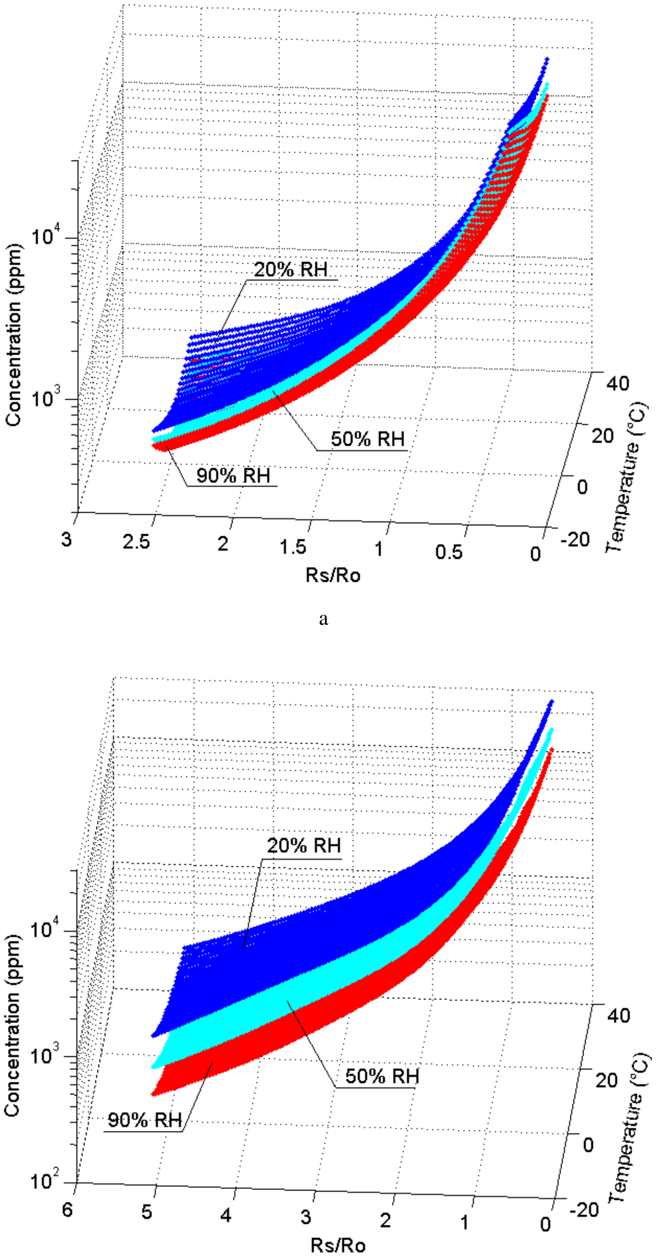
Figure 5

ANN for approximating the gas sensor characteristics

Fig. 6 shows the results from the output of trained neural networks for sensors TGS813 and TGS2611 and surfaces with different humidity levels are illustrated.

Thus, on the basis of three-dimensional approximation of sensor characteristics resulting from ANN training, the value of methane concentration can be obtained when the impact of temperature and relative humidity is compensated for.

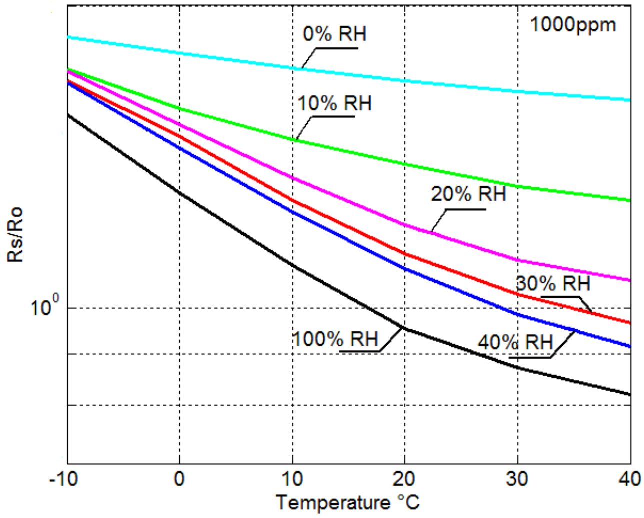
Fig. 7 shows the characteristics obtained by ANN, illustrating the joint impact of ambient temperature and humidity on the sensors resistance ratio, Rs/Ro, at a concentration of 1000ppm for sensor TGS813 and 5000ppm for sensor TGS2611.



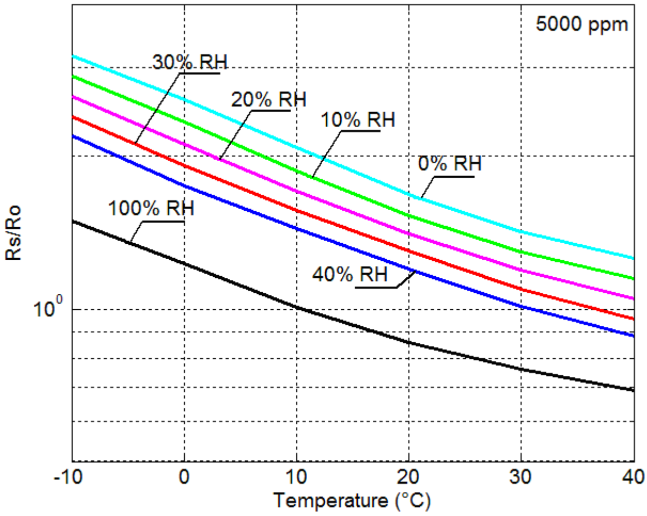
a

b

Figure 6
Results from the output of the trained neural network: a) for sensor TGS813; b) for sensor TGS2611



a



b

Figure 7

Impact of ambient temperature and humidity on sensors resistance ratio Rs/Ro: a) of sensor TGS813 at referent concentration of 1000ppm; b) of sensor TGS2611 at referent concentration of 5000ppm

The algorithm for compensating for the impact of temperature and humidity on gas sensors readings by means of ANN in the process of gas control is shown in Fig. 8.

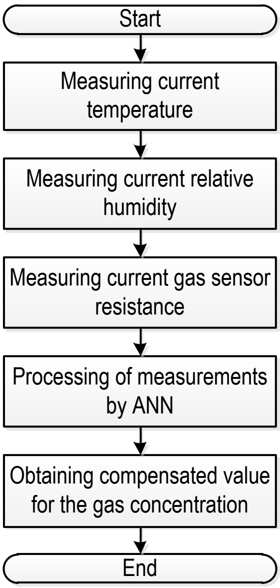

Start
Measuring current temperature
Measuring current relative humidity
Measuring current gas sensor resistance
Processing of measurements by ANN
Obtaining compensated value for the gas concentration
End

Figure 8
ANN compensation algorithm in gas sensors

To estimate the error which occurs if there is no compensation for temperature and humidity on sensor characteristics, the absolute error

$$\Delta Conc_{t,RH} = Conc - Conc_{t,RH} \tag{7}$$

and normalized error

$$\varepsilon_{n\ t,RH} = \frac{\Delta Conc_{\ t,RH}}{Conc_{\max} - Conc_{\min}}.100\% , \tag{8}$$

are calculated, where *Conc* is the concentration based on the operating characteristic without compensation; $Conc_{t,RH}$ is the real concentration corresponding to characteristics given the variation in temperature and relative humidity; and $Conc_{\max} - Conc_{\min}$ is the range of concentration variation for each sensor.

The errors occurring for 1000ppm and 2000 ppm when using the basic characteristics of sensors at 20°C/65%RH, without taking into account the variation in temperature and in relative humidity, are shown in Tables 1 and 2 respectively.

Table 1

Normalized error for sensor TGS813 without compensation when using the basic characteristic at 20°C/65%RH

t°C /%RH	*Conc*, ppm	$\varepsilon_{n\ t,RH}$, %
-10°C / 0%RH	1000	-28.7
-10°C / 0%RH	2000	-60.0
40°C/100%RH	1000	3.7
40°C/100%RH	2000	7.3

Table 2

Normalized error for sensor TGS2611 without compensation when using the basic characteristic at 20°C/65%RH

t°C /%RH	*Conc*, ppm	$\varepsilon_{n\ t,RH}$, %
-10°C / 0%RH	1000	-16.5
-10°C / 0%RH	2000	-32.5
40°C/100%RH	1000	5.9
40°C/100%RH	2000	11.9

These results confirm the necessity of compensating for the impact of temperature and relative humidity.

Using the trained neural network, the values $Conc_{ANN}$ of methane concentration have been obtained at various values of resistance, temperature and humidity. The absolute error is determined based on these values

$$\Delta Conc = Conc_{ANN} - Conc \tag{9}$$

and the normalized error of the ANN method is

$$\varepsilon_n = \frac{\Delta Conc}{Conc_{\max} - Conc_{\min}}.100\%\ , \tag{10}$$

where $Conc_{ANN}$ is the concentration value, determined using the trained neural network; $Conc$ is the respective real concentration value from the basic experimental characteristics which have not taken part in training; and $Conc_{\max} - Conc_{\min}$ is the range of concentration variation for each sensor.

The experimental basic characteristics, which have not been used in training, and those obtained by ANN for the two sensors are shown in Fig. 9.

Fig. 10 gives a graphic presentation of normalized errors (equation (10)) when employing the proposed method for compensation by ANN.

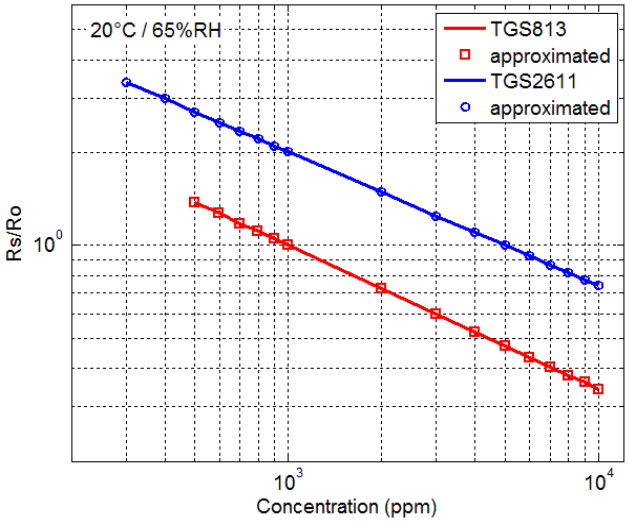


Figure 9

Experimental basic characteristics and characteristics obtained by ANN

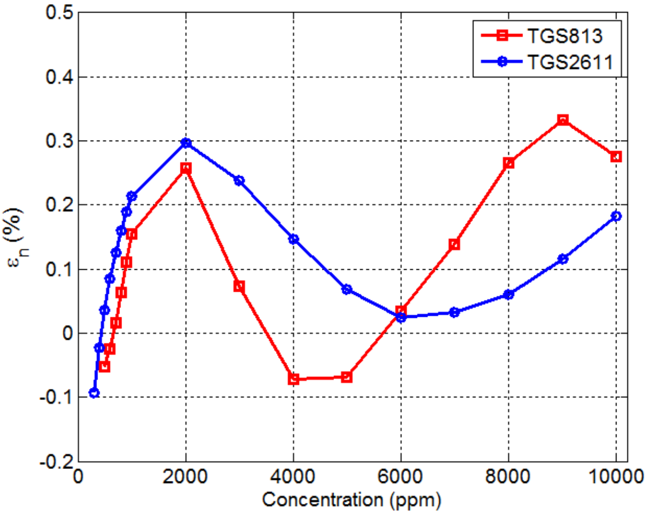


Figure 10

Normalized errors when implementing the ANN-compensation method

The obtained results show that the normalized error of the ANN method is in the range of -0.05% to +0.35% for sensor TGS813, and -0.1% to +0.3% for sensor TGS2611, which confirms the effectiveness of the implementation of the proposed ANN compensation method.

Conclusions

On the basis of the research conducted, the following conclusions can be drawn:

- ambient temperature and humidity have a substantial impact on metal oxide gas sensor characteristics;
- a method is proposed for compensating for the impact of temperature and humidity on gas sensors by using ANN for three-dimensional approximation of their characteristics;
- to compensate for the impact of ambient conditions on gas sensors of type TGS813 and TGS261 for the measurement and control of methane, a trained ANN with back propagation of error with two hidden (input and intermediate) and one output layers has been obtained;
- the three-dimensional approximation of gas sensor characteristics by the trained neural network allows the impact of temperature and humidity to be compensated for and the normalized error is from -0.05% to +0.35% for sensor TGS813 and from -0.1% to +0.3% for sensor TGS2611.
- the proposed method of compensating for the impact of disturbing factors by ANN can also be used for other types of sensors, as well as for performing higher order approximation with a greater number of disturbing factors.

References

[1] G. F. Fine, L. M. Cavanagh, A. Afonja and R. Binions: Metal Oxide Semi-Conductor Gas Sensors in Environmental Monitoring, Sensors 2010, 10, pp. 5469-5502

[2] M. Fleischer, M. Lehmann: Solid State Gas Sensors - Industrial Application, Springer-Verlag Berlin Heidelberg, 2012, p. 269

[3] K. Arshak, E. Moore, G. M. Lyons, J. Harris and S. Clifford: A Review of Gas Sensors Employed in Electronic Nose Applications, Sensor Review, Vol. 24, Number 2, 2004, pp. 181-198

[4] A. D. Wilson and M. Baietto: Applications and Advances in Electronic-Nose Technologies, Sensors, 2009, 9, pp. 5099-5148

[5] S. Zampolli, I. Elmi, F. Ahmed1, M. Passini, G. C. Cardinali, S. Nicoletti, L. Dori: An Electronic Nose Based on Solid State Sensor Arrays for Low-Cost Indoor Air Quality Monitoring Applications, Sensors and Actuators B 101, 2004, pp. 39-46

[6] T. Nenov, P. Panteleev: Gas Sensors for Environmental Monitoring. Automatica&Informatics, 2010, No. 1, pp. 16-19

[7] FIGARO Engineering Inc. Products - Gas Sensors (www.figaro.co.jp/en/product/)

[8] SYNKERA Technologies Inc. Products (www.synkera.com)

[9] e2v Technologies. Products (www.e2v.com)

[10] Sencera. Products (www.sencera.com)

[11] N. Barsan and U. Weimar: Understanding the Fundamental Principles of Metal Oxide-based Gas Sensors; the Example of CO Sensing with SnO_2 Sensors in the Presence of Humidity, J. Phys.: Condens. Matter, 2003, 15, pp. 813-839

[12] K. Shimizu, I. Chinzei, H. Nishiyama, S. Kakimoto, S. Sugaya, W. Matsutani, A. Satsuma: Doped-Vanadium Oxides as Sensing Materials for High Temperature Operative Selective Ammonia Gas Sensors. Sensors and Actuators B 141, 2009, pp. 410-416

[13] N. Han, L. Chai, Q. Wang, Y. Tian, P. Deng, Y. Chen: Evaluating the Doping Effect of Fe, Ti and Sn on Gas Sensing Property of ZnO, Sensors and Actuators B 147, 2010, pp. 525-530

[14] Ch.-Y. Lin, Y.-Y. Fang, Ch.-W. Lin, J. J. Tunney, K.-Ch. Ho: Fabrication of NO_x Gas Sensors Using In_2O_3–ZnO Composite Films, Sensors and Actuators B 146, 2010, pp. 28-34

[15] Z. Jiang, Z. Guo, B. Sun, Y. Jia, M. Li, J. Liu: Highly Sensitive and Selective Butanone Sensors Based on Cerium-doped SnO_2 Thin Films, Sensors and Actuators B 145, 2010, pp. 667-673

[16] G. Korotcenkov, B. K. Cho: Thin Film SnO_2-based Gas Sensors: Film Thickness Influence, Sensors and Actuators B 142, 2009, pp. 321-330

[17] Ch. Wang, L. Yin, L. Zhang, D. Xiang and R. Gao: Metal Oxide Gas Sensors: Sensitivity and Influencing Factors, Sensors 2010, 10, pp. 2088-2106

[18] J. Janiczek: Approximation of a Multivariable Sensor Characteristic Applied to the Carbon Monoxide Sensor, Measurement, 2010, 43, pp. 1115-1118

[19] E. L. Hines, J. W. Gardner: An Artificial Neural Emulator for an Odor Sensor Array, Sensors and Actuators B, 18-19, 1994, pp. 661-664

[20] W. Ping, X. Jun: A Novel Recognition Method for Electronic Nose Using Artificial Neural Network and Fuzzy Recognition, Sensors and Actuators B 37, 1996, pp. 169-174

[21] H.-K. Hong, Ch. H. Kwon, S.-R. Kim, D. H. Yun, K. Lee, Y. K. Sung: Portable Electronic Nose System with Gas Sensor Array and Artificial Neural Network, Sensors and Actuators B 66, 2000, pp. 49-52

[22] D. Luo, H. G. Hosseini, J. R. Stewart: Application of ANN with Extracted Parameters from an Electronic Nose in Cigarette Brand Identification, Sensors and Actuators B 99, 2004, pp. 253-257

[23] S. Osowski, T. H. Linh, K. Brudzewski: Neuro-Fuzzy Network for Flavor Recognition and Classification, IEEE Transactions on Instrumentation and Measurement, Vol. 53, No. 3, June 2004, pp. 638-644

[24] A. P. Singh, S. Kumar, T. S. Kamal: Virtual Compensator for Correcting the Disturbing Variable Effect in Transducers, Sensors and Actuators A 116, 2004, pp. 1-9

[25] T. Nenov, S. Ivanov: Linearization of Characteristic of Relative Humidity Sensor and Compensation of Temperature Impact, Sensors and Materials, 2007, Vol. 18, No. 2, pp. 95-106

[26] W. Hao, X. Li, M. Zhang: Application of RBF Neural Network to Temperature Compensation of Gas Sensor, Proceeding CSSE '08 Proceedings of the 2008 International Conference on Computer Science and Software Engineering - CSSE '08, Vol. 04, pp. 839-842

Sensor-based Navigation and Integrated Control of Ambient Intelligent Wheeled Robots with Tire-Ground Interaction Uncertainties

Aleksandar Rodic[1], Gyula Mester[2]

[1] University of Belgrade, Institute Mihajlo Pupin, Robotics Laboratory
Volgina 15, 11060 Belgrade, Serbia, aleksandar.rodic@pupin.rs

[2] University of Szeged, Faculty of Engineering, Robotics Laboratory
Mars tér 7, 6724 Szeged, Hungary, gmester@inf.u-szeged.hu

Abstract: This paper regards the synthesis of intelligent non-visual sensor-based navigation, motion planning and the integrated control of indoor ambient adaptive wheel-based mobile robots in unknown environments with tire-ground interaction uncertainties. The problem relates to searching appropriate techniques how to navigate towards a target position in an unknown environment when the obstacles to avoid are discovered in real time, and how to maintain collision free motion of a high dynamic performance. Environments characterized by variable ground surface conditions with immobile obstacles of different shapes and sizes will be considered in the paper as unexpected disturbances, i.e. system uncertainties. The tools developed to address this issue thus consist of the combination of cognitive motion planning and control theory techniques, including a non-linear model-based approach. Two characteristic approaches to integrated control are evaluated in the paper: a kinematical as well as dynamic one, in the sense of control efficiency and robustness to the environmental and model uncertainties. Characteristic simulation tests are performed to verify the proposed algorithms.

Keywords: mobile robots; sensor-based navigation; integrated control; tire-ground interaction

1 Introduction

Mobile wheel-based robots are subjected to many recent research studies with aim to provide reliable and robust robotic platforms for broad service applications at home, in office, and at public institutions. Mobile robotic platforms form the basis for the building (development) of high-tech devices, such as personal robots of high performance to be widely used in the future in everyday human life. The problem of the use of such advanced intelligent systems is related to the success in solving complex cognitive and control tasks, such as intelligent navigation, motion planning and robust control of the system dynamics in conditions of unknown, unpredicted and evolving environments. The problem of ambient adaptation to the unstructured, confined and cluttered environments as well adaptation to the variable ground surface conditions (contingency risks) requires the development of efficient control techniques. A combined knowledge-based and model-based algorithm that couples navigation and control capabilities into a unified control architecture designed for the accurate system navigation and integrated control of wheeled robot dynamics is proposed in the paper. Autonomous mobile robots are required to have high dynamic performance, in the sense of dynamic, non-jerky and smooth motion, in order to ensure the reliable performance of tasks imposed.

Numerous research studies concerning control of wheeled mobile robots were reported in [1], [2]. In particular, non-holonomy constraints associated with these systems have motivated the development of highly nonlinear control techniques. For the sake of simplicity, the control methods are developed mainly for car-like mobile robots. The heuristic methods were the first techniques used to generate motion based sensors. The majority of these works were derived from classic planning methods [3]. The methods of physical analogies assimilate obstacle avoidance to a known physical problem. The representative of them is the potential field method [4]. There are methods that compute some high-level information as intermediate information, which is translated next in motion. The nearness diagram navigation [5, 6] is a representative of this method. In paper [7] a multi-agent, self-organizing system of mobile robots is controlled by the implementation of the genetic algorithm. The solution proposed in paper [8] represents an original approach to the design of the 2-DOF Takagi–Sugeno PI-fuzzy controller based on the stability analysis theorem. For implementation of the ethologically inspired robot behavior, a platform based on a fuzzy state-machine was suggested in [9]. In paper [10], the method of utilization of the low-resolution data for control purposes was applied. In this paper, control is based on fuzzy logic, with the deployment of stochastic digital low-resolution time arrays. The imprecision of the control method was eliminated by stochastic noise superimposed during data gathering, while the negative effects of noise are suppressed both by the fuzzy nature of the decision-making process and by the energy inertia of the controlled object.

The considerations to be conducted in the paper will demonstrate a methodology of the motion control of mobile robots that combines ad-hoc motion planning and obstacle avoidance together with model-based, integrated control of the robotic system. The paper is organized as follows: Section 1: introduction. In Section 2, the modeling of wheel-based robots is presented. In Section 3, sensor-based navigation and motion planning are illustrated. In Section 4, the motion control is presented. Simulation examples and verification of the proposed methodology are illustrated in Section 5. Conclusions are given in Section 6.

2 Modeling of Wheel-based Robots

For the purpose of control system development and the simulation model of a non-holonomic wheeled robot with differential (skid) steering is considered in the paper as presented in Fig. 1. A 2WD indoor mobile robotic platform with differential steering and two auxiliary (passive, i.e. non-powered) wheels is assumed in the paper as a system representative. A non-linear model of such a wheeled mobile robot is considered, taking into account that the robot can move on a sloped surface, too. In the general case, surface inclination angle can appear in both longitudinal γ_x as well as lateral γ_y direction of motion (Fig. 2) with respect to the longitudinal x-axis of the robot body. The direction of the motion, i.e. the angle of the forward (transport) speed vector $\vec{\mathbf{V}}$ (Fig. 1), depends on amplitudes of particular tire angular velocities, rigid-body parameters as well as tire-ground interaction parameters and ground surface condition. The referent coordinate system $OXYZ$ to be considered in the paper is attached to the ground surface. The local mobile coordinate system $oxyz$ is attached to the mass center (MC) of the wheeled mobile robot. Robot motion is consequence of differential steering, i.e. controlled changing of tires r.p.m. corresponding longitudinal $F_{xi}, i=1,2$ and lateral $F_{yi}, i=1,2$ tire forces (Fig. 2), which cause the robot to move in the desired direction and with the desired forward speed. The passive auxiliary tires have zero traction forces $F_{xi}=0, i=3,4$. In the general case, when the auxiliary wheels (i.e. their axles) are not collinear with the active (powered) wheels, they can slide on the surface against the friction forces. In this case, the lateral tire forces $F_{yi}\neq 0, i=3,4$ appear on the auxiliary wheels. In the general case, the forward speed $\vec{\mathbf{V}}$ is not collinear with the direction of the longitudinal wheeled mobile robot axis of symmetry. The angle between the velocity vector $\vec{\mathbf{V}}$ and the longitudinal x-axis of symmetry is defined by the angle β known as the slip angle of vehicle [11]. The particular mobile robot wheels perform corresponding rotational as well linear movements. Linear tire velocities are signed by $\vec{v}_i, i=1,\ldots,4$ in Fig. 1. In the general case, these velocities do not coincide with the corresponding direction of robot motion defined by the vector $\vec{\mathbf{V}}$. The consequence of this is the appearance of tire slipping defined by the corresponding angles $\varsigma_i, i=1,\ldots,4$ as presented in Fig. 1. Some important geometry parameters of the wheeled robot (rover) are presented in Fig. 1. These are: b is the track of rover, l_x is the relative position of the active wheels with respect to the robot mass centre (MC) observed along the longitudinal *x*-coordinate direction, l_f and l_r are corresponding distances of the auxilary wheels (front and rear) from the robot MC.

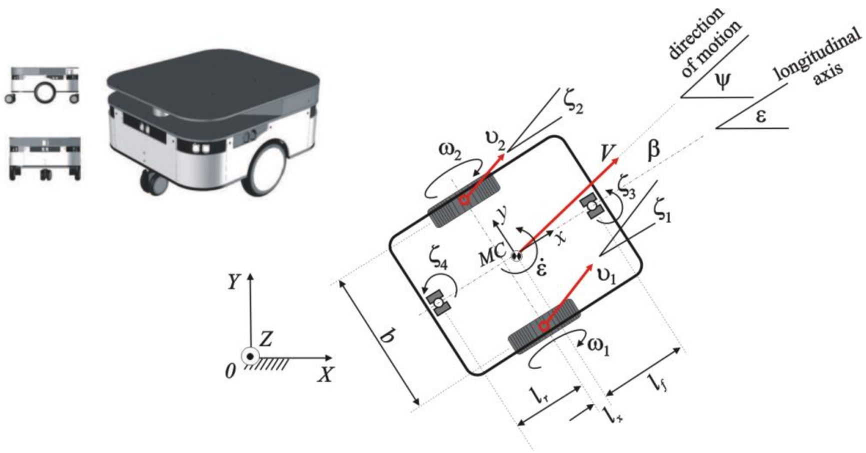


Figure 1
Industrial non-holonomic 2WD wheel-based mobile robot RobuLab10 [12]. Corresponding kinematic model of the assumed robotic platform and parameters of interest.

The vector of the state variables, expressed with respect to the referent coordinate system *OXYZ* can be written in the form:

$$\mathbf{q} = [X \; Y \; \varepsilon]^T \tag{1}$$

where *X* and *Y* represent corresponding linear displacements (translations) of robot body MC determined in the absolute coordinate system attached to the ground surface, ε is corresponding yaw angle, i.e. turning of the rover about the Z-axis measured with respect to the X-axis (Fig. 1).

2.1 Kinematical Model of a 2WD Robot

The kinematical model of the robot presented in Fig. 1, with two active and two auxiliary non-powered wheels can be defined by the following relation:

$$\begin{bmatrix} V \\ \dot{\varepsilon} \end{bmatrix} = \begin{bmatrix} r_t/2 & r_t/2 \\ r_t/b & -r_t/b \end{bmatrix} \cdot \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} \tag{2}$$

where V and $\dot{\varepsilon}$ represent corresponding amplitudes of the forward (transport) speed as well as yaw-rate of robot rigid-body, r_t is tire radius, and ω_1 and ω_2 are corresponding right and left tire angular velocities. The transport speed $\vec{\mathbf{V}}$, determined in the plane of motion, has two components – the longitudinal $\dot{x}$ and lateral $\dot{y}$:

$$V = \left\| \vec{\mathbf{V}} \right\| = \sqrt{\dot{x}^2 + \dot{y}^2} \tag{3}$$

Two kinematical variables are needed for tire modeling: tire slip ratio s_i and tire slip angle α_i. These variables can be calculated [11, 12] for every particular robot tire using the following relations:

$$s_i = \frac{v_i \cos(\alpha_i) - r_i\,\omega_i}{r_i\,\omega_i},\ i = 1,\ldots,4 \tag{4}$$

$$\alpha_i = -\zeta_i,\ \ i = 1,\ldots,4 \tag{5}$$

The powered tires i=1,2 have no possibility of steering in a direct way while the passive wheels i=3,4 are free for turning about the vertical axis to enable better system maneuverability (Fig. 1). The value v_i is the corresponding linear speed of the centre of mass of the particular *i*-th robot tire, and ζ_i represents the so-called tire speed angle defined with respect to the longitudinal direction of motion collinear to the longitudinal x-axis of symmetry. The translational (linear) tire speeds of tire mass centers are determined by the relations:

$$\begin{aligned} v_1 &= \sqrt{(\dot{y} - l_x\dot{\varepsilon})^2 + (\dot{x} - b/2\,\dot{\varepsilon})^2} \quad \text{for right wheel} \\ v_2 &= \sqrt{(\dot{y} - l_x\dot{\varepsilon})^2 + (\dot{x} + b/2\,\dot{\varepsilon})^2} \quad \text{for left wheel} \\ v_3 &= \sqrt{(\dot{y} + l_f\dot{\varepsilon})^2 + \dot{x}^2} \quad \text{for front wheel} \\ v_4 &= \sqrt{(\dot{y} - l_r\dot{\varepsilon})^2 + \dot{x}^2} \quad \text{for rear wheel} \end{aligned} \tag{6}$$

where the directions of the particular tire velocities (6) are determined by the corresponding angles ζ_i, which are calculated from the expressions:

$$\begin{aligned} tg(\zeta_1) &= \frac{\dot{y} - l_x\dot{\varepsilon}}{\dot{x} - b/2\,\dot{\varepsilon}}, \quad tg(\zeta_2) = \frac{\dot{y} - l_x\dot{\varepsilon}}{\dot{x} + b/2\,\dot{\varepsilon}} \\ tg(\zeta_3) &= \frac{\dot{y} + l_f\dot{\varepsilon}}{\dot{x}}, \quad tg(\zeta_4) = \frac{\dot{y} - l_r\dot{\varepsilon}}{\dot{x}} \end{aligned} \tag{7}$$

2.2 Model of 2WD Robot Rigid-Body Dynamics

The model of rigid-body dynamics of the assumed 2WD mobile robot (Fig. 1) is presented in Fig. 2. The dynamic model of this robotic system can be defined in the following form:

$$\mathbf{T} = \mathbf{H}(\mathbf{q}) \cdot \ddot{\mathbf{q}} + \mathbf{h}_{\mathbf{ccg}}(\mathbf{q}, \dot{\mathbf{q}}) - \mathbf{F}_{\mathbf{w}}(\dot{\mathbf{q}}) - \mathbf{J}_{\mathbf{s}} \cdot \mathbf{S} \tag{8}$$

where $\mathbf{T} \in \Re^{3\times1}$ is a vector of the generalized forces/torques acting in robot MC. The vector $\mathbf{T}$ has three components collinear to the main coordinate directions *X*, *Y* and *Z* (Fig. 2): the generalized forces T_X and T_Y and corresponding generalized torque T_ε about z-axis; $\mathbf{H} \in \Re^{3\times3}$ is an inertia matrix of the robot-body; $\mathbf{h}_{\mathbf{ccg}} \in \Re^{3\times1}$ is a vector of centrifugal, Coriolis and gravity forces; $\mathbf{F}_{\mathbf{w}} \in \Re^{3\times1}$ is a vector of external resistance forces and torques that includes aerodynamic resistance, rolling resistance and Coulomb friction forces. Body impact forces (moments) as the consequence of an occasional strike of the robot to the surrounding objects are also taken into account by the vector $\mathbf{S} \in \Re^{3\times1}$. Corresponding Jacobian is defined by $\mathbf{J}_{\mathbf{S}} \in \Re^{3\times3}$. The vector of generalized forces **T**, defined by (8), can be expressed with respect to the mobile coordinate system MC-*xy*. Then, it can be defined in the form:

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_x \\ \tau_y \\ \tau_\varepsilon \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{4} F_{xi} \\ \sum_{i=1}^{4} F_{yi} \\ M_Z \end{bmatrix}; \quad F_{x3} = 0; \quad F_{x4} = 0 \tag{9}$$

$$M_Z = \left(F_{x1} - F_{x2}\right) \cdot \frac{b}{2} + \left(F_{y1} + F_{y2}\right) \cdot l_x + F_{y3} \cdot l_f - F_{y4} \cdot l_r \tag{10}$$

where F_{x_i} and F_{y_i} (i=1,..,4) are the corresponding longitudinal and lateral traction (braking) tire forces (Fig. 2) while *b*, l_x, l_f and l_r are the constructive parameters presented in Fig. 1.

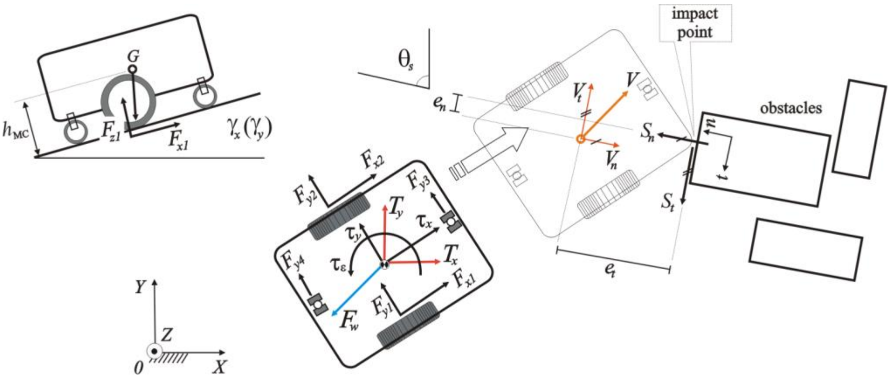


Figure 2

Dynamic model of the assumed 2WD wheel-based robot including traction (braking) forces, resistance forces, side impact forces and slope effects

The relation between the generalized forces and torques **T** expressed in the absolute coordinate system $0XYZ$ and corresponding forces and torques $\boldsymbol{\tau}$ defined in the local coordinate system MC-*xyz* can be defined in the following way:

$$\begin{aligned} T_X &= \tau_x \cos(\varepsilon) - \tau_y \sin(\varepsilon), \\ T_Y &= \tau_x \sin(\varepsilon) + \tau_y \cos(\varepsilon), \\ T_\varepsilon &= \tau_\varepsilon \end{aligned} \tag{11}$$

The corresponding matrix and vectors given in (8) are assumed in the form [11]:

$$\mathbf{H} = \begin{bmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & I_Z \end{bmatrix} \tag{12}$$

$$\mathbf{h_{ccg}} = \begin{bmatrix} -m\,\dot{y}\,\dot{\varepsilon} + m\,g\sin(\gamma_x) \\ m\,\dot{x}\,\dot{\varepsilon} + m\,g\sin(\gamma_y) \\ 0 \end{bmatrix} \tag{13}$$

where m is the lump mass of robot system, I_z is the robot’s axial moment of inertia with respect to the z- axis, and g is the magnitude of gravity acceleration. The resultant vector of the aerodynamic resistance as well as the rolling resistance forces and torques is calculated in a way [11]:

$$\mathbf{F_w} = \begin{bmatrix} -K_x\dot{x}^2 - \sum_{i=1}^{4} f_{r_i} F_{z_i} \cos\zeta_i \\ -K_y\dot{y}^2 - \sum_{i=1}^{4} f_{r_i} F_{z_i} \sin\zeta_i \\ M_\alpha \\ \end{bmatrix} \tag{14}$$

where $\mathrm{K_x}$, $\mathrm{K_y}$ represents corresponding air resistance coefficients of robot body; M_α is a sum of tire self-aligning torques (slipping resistances about the vertical axis) due to robot turning about vertical axis; f_{r_i} is a rolling resistance coefficient of the i-th tire, and F_{z_i} represents corresponding tire payload.

The impact force vector $\mathbf{S}$ has in a general case three particular components $\mathbf{S} = [S_n\ S_t\ S_\varepsilon]^T$ (Fig. 2). The variation of motion quantity exchanged during the robot strike in the particular impact point (Fig. 2) is equal to the impulse of the impact force S_n produced in the direction $\vec{n}$. The tangential impact force component S_t depends on strike magnitude and corresponding body friction coefficient. Bearing in mind what has been previously said, the following relations can be derived:

$$\mathbf{S_n} = \frac{m \cdot V \cdot \cos(\theta_s)}{\Delta t} \cdot \vec{\mathbf{n}} \tag{15}$$

where Δt represents the time interval of impact impulse. The tangential component of impact force S_t is calculated from the relation:

$$\mathbf{S_t} = -\mu_s \cdot S_n \cdot \vec{\mathbf{t}} \tag{16}$$

where μ_s is Coulomb's friction coefficient characteristic for a relative two bodies sliding, i.e. robot-object interaction. In the general case, an impact force causes additional rotation of the robot body due to the particular location e_n of the impact point with respect to the MC (Fig. 2). Then, the turning moment of the impact force about the axis that passes through the MC can be defined in a way, taking into account that $\vec{\mathbf{p}} = \vec{\mathbf{n}} \times \vec{\mathbf{t}}$:

$$\mathbf{S_\varepsilon} = \left(S_n \cdot e_n - S_t \cdot e_t\right) \cdot \vec{\mathbf{p}} \tag{17}$$

Jacobian $\mathbf{J_s}$ in (8) is determined in a form of the transformation matrix that is calculated for the case of system rotation about the z-axis for the angle θ_s as the relative angle between local $x - y$ and $n - t$ coordinate systems (see Fig. 2). In this case, the impact forces and torques $\mathbf{J_s} \cdot \mathbf{S}$ in (8) can be calculated by the relation:

$$\mathbf{J_s} \cdot \mathbf{S} = \begin{bmatrix} -\sin(\theta_s) & \cos(\theta_s) & 0 \\ \sin(\theta_s) & \cos(\theta_s) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} S_n \\ S_t \\ S_\varepsilon \end{bmatrix} \tag{18}$$

2.3 Non-linear Tire Model

A considerable number of different models of tire force and moment generating properties have been proposed in the available literature. The standard description of vehicle tire dynamics is the so called the magic formula tire model originally introduced by Pacejka and Bakker [13]. The model provides a set of mathematical formulae from which the forces and moment acting from road to tire can be calculated at longitudinal and lateral slip conditions, which may occur simultaneously. The formula (model) expresses the side force F_y, the aligning torque M_α and the longitudinal force F_x as a function of two arguments - the side slip angle defined by (5) and the longitudinal tire slip ratio determined by (4), respectively. The general form of the formula, which holds for a given value of vertical tire load, looks like:

$$f(u) = D \cdot \sin\{C \cdot arctg[B \cdot u - E \cdot (B \cdot u - arctg(B \cdot u))]\} \tag{19}$$

The empiric non-linear Pacejka's tire model is shown in Fig. 3. For constant coefficients B, C, D and E the curve exhibits an anti-symmetric shape with respect to the origin. The formula is capable of producing characteristics which closely match measured curves for longitudinal F_x, side (lateral) F_y force and self-aligning torque M_α as functions of their respective slip quantities: the slip angle α and longitudinal slip ratio s. The output variable stands for either F_x, F_y or M_α and the input u may represents s or α.

Tire payload, i.e. tire torques of rotation, can be now calculated from the relation:

$$\tau_t = F_x \cdot r_t \tag{20}$$

taking into account the longitudinal tire force F_x and tire radius r_t. Tire payload τ_t as well as tire angular velocity are used as feedback signals for servo-control of wheeled robot motion.

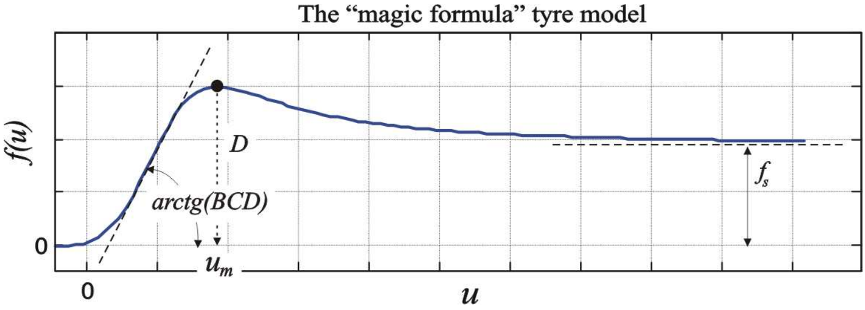


Figure 3
The Magic formula tire model – a graphic presentation

The model of wheeled robot determined by relations (2 - 20) is used in the chapter for synthesis of algorithms for control of robot dynamics in the case of variation of tire-ground adhesion parameters, i.e. in the presence of tire-ground interaction uncertainties.

3 Sensor-based Navigation and Motion Planning

The objective of motion planning techniques is to compute a collision-free trajectory to the target configuration that complies with the vehicle constraints (Fig. 4). The objective is to move a vehicle towards a target location free of collisions with the obstacles detected by the sensors during motion execution. The advantage of reactive obstacle avoidance is to compute motion by introducing the sensor information within the control loop, used to adapt the motion to any contingency incompatible with the initial plans.

The approach that is elaborated in this section assumes that the robot is equipped with the corresponding sensors for the detection of obstacles (sonar proximity sensors and range-finder sensors) in its surrounding, as well as that it operates in informatically structured environments [14]. This assumes it should be connected to a wireless sensor network that ensures accurate localization of robot and target point (Fig. 4) within the work-space at every time instant. When autonomous robot moves towards the target position and its sensors detect obstacle(s), a corresponding avoiding strategy should be activated. In that sense, robot motion can be described as a compromise between avoiding obstacles and moving towards the target position as presented in Fig. 4. Autonomous robots react to both of the sensed variables (target direction and collision free direction of motion with respect to neighbourhood obstacles) to perform autonomous motion.

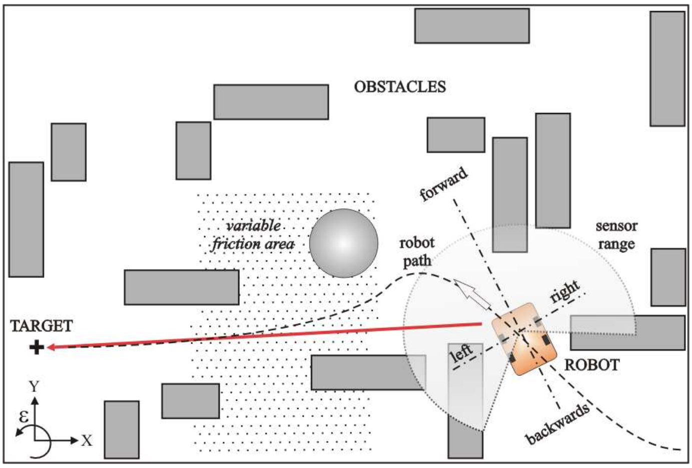


Figure 4

Model of sensor-based navigation and motion planning in presence of obstacles with variable tire-ground interaction conditions

For accurate navigation and motion planning in the presence of obstacles, the robot needs to sense the following variables (see Fig. 4): (i) distance to surrounding obstacle(s) within the sensor range (front-rear and right-left sectors of acquiring information), (ii) corresponding relative position of obstacles (i.e. angles measured relatively to the robot longitudinal x-axis of symmetry), and (iii) relative position (azimuth angle) of robot with respect to the target point. Moving towards the target point and avoiding obstacles along the path predicted (Fig. 4), a mobile robot changes its orientation in the work-space as well as its forward velocity. When an obstacle is detected by sensors, the mobile robot slows down and changes its direction of motion according to the actual conditions detected. The navigation strategy of mobile robot is set to enable the automatic guidance of robot in the presence of obstacles and the accurate tracking of the estimated target direction. For the purpose of spatial reasoning in unknown and confined environments, an appropriate fuzzy inference system (FIS) is commonly used technique [14-17] to support such kind of cognitive tasks. In this paper, the FIS system presented in Fig. 5 is designed for collision free robot guidance in unknown environments with the presence of obstacles of different shapes and sizes.

The fuzzy navigation system, assumed in the paper to navigate robot in unknown, informatically structured environments, is shown in Fig. 5. It represents a multi-input / multi-output (MIMO) system with six inputs and two output ports.

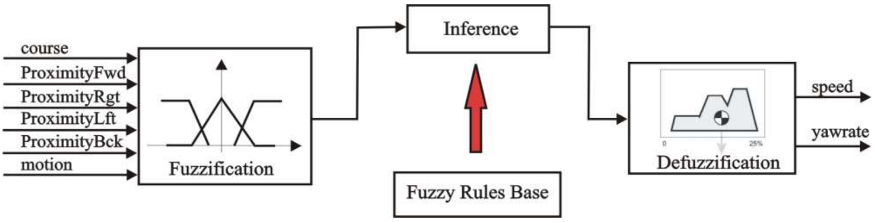


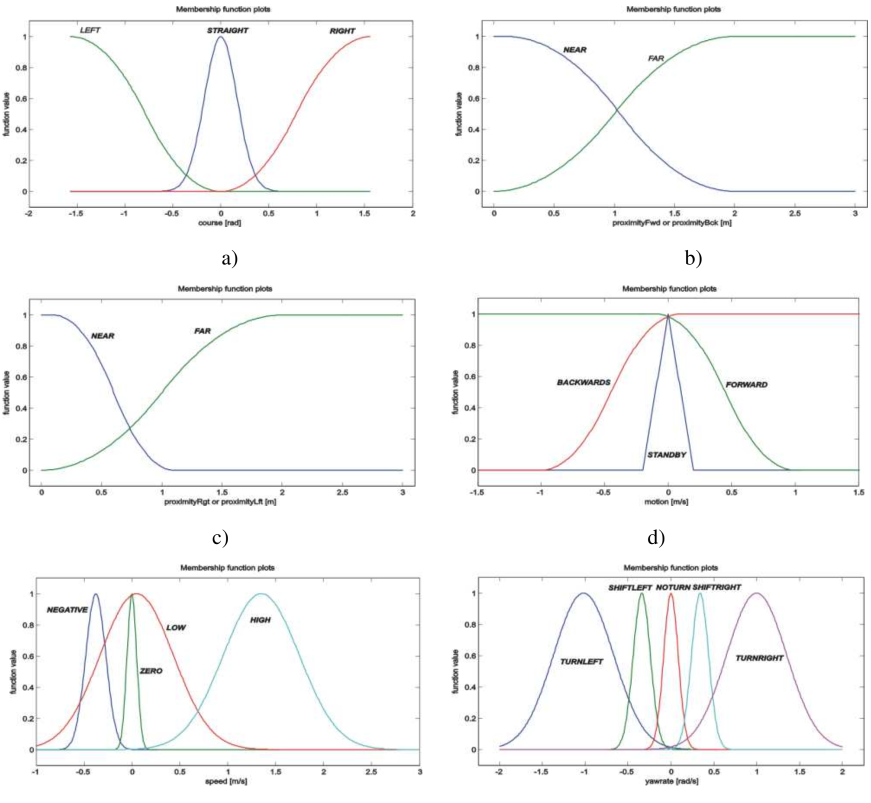
Figure 5

Block scheme of the Fuzzy Inference System developed for sensor-based navigation and obstacle avoidance in unknown environments

The following input variables are needed to be acquired from the work-space: target direction (azimuth angle of the referent course towards to the goal position); the “course” angle is determined relatively to the local coordinate system attached to the robot MC, with x-axis oriented along the longitudinal axis of robot platform (Fig. 1). The course is positive when the robot has to turn in left and negative when it moves right; proximity (distance) to obstacles in the forward direction; side proximity to obstacles (right-hand side and left-hand side direction); proximity to obstacles with respect to backwards direction; and indicator of motion concerning movement forward, backwards or standby status.

The corresponding output variables of the FIS considered in Fig. 5 are forward speed and yaw-rate of the wheeled robot. The FIS membership functions (MF) that correspond to the particular fuzzy input ports are presented in Figs. 6a-6d. The corresponding MF that corresponds to the output fuzzy ports are shown in Figs. 6e and 6f. The MFs chosen in this robot task have predominantly Gauss form.

The FIS rule database consists of 13 rules that allow the system to navigate properly in the presence of immobile obstacles. The rule data-base is systematized in Tab. 1. Weighting factors related to the particular rules are set to the unit value (Tab. 1). Membership functions and fuzzy rules are designed in the Fuzzy Logic Toolbox of Matlab/Simulink.



a) b)

c) d)

e) f)

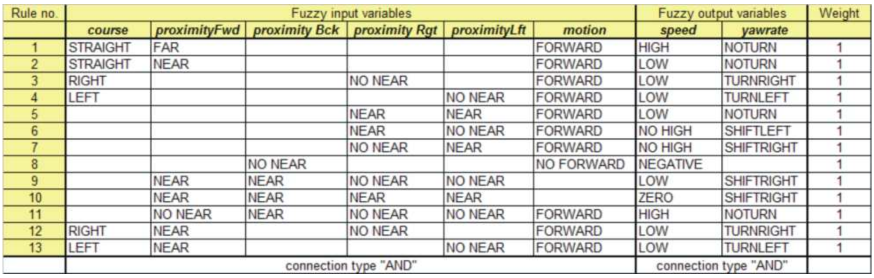
Figure 6

Fuzzy membership functions used for sensor-based navigation

input variable: a) "course " [-1.57, 1.57 rad], b) distance from the obstacles in forward/backwards direction, "proximityFwd" or "proximityBck", [0,3 m], c) distance to obstacles in side directions – right and left,"proximityRgt or proximityLft", [0, 3 m], d) status of motion, "motion", [-1.5, 1.5 m/s], output variable: e) forward speed "speed", [-1, 3 m/s], f) yaw-rate "yawrate" [-2, 2 rad/s].

Table 1

Data-base fuzzy rules designed to ensure robot navigation in unknown environments with obstacles of different geometry and size



Rule no.	Fuzzy input variables						Fuzzy output variables		Weight
	course	*proximityFwd*	*proximity Bck*	*proximity Rgt*	*proximityLft*	*motion*	*speed*	*yawrate*	
1	STRAIGHT	FAR				FORWARD	HIGH	NOTURN	1
2	STRAIGHT	NEAR				FORWARD	LOW	NOTURN	1
3	RIGHT			NO NEAR		FORWARD	LOW	TURNRIGHT	1
4	LEFT				NO NEAR	FORWARD	LOW	TURNLEFT	1
5				NEAR	NEAR	FORWARD	LOW	NOTURN	1
6				NEAR	NO NEAR	FORWARD	NO HIGH	SHIFTLEFT	1
7				NO NEAR	NEAR	FORWARD	NO HIGH	SHIFTRIGHT	1
8			NO NEAR			NO FORWARD	NEGATIVE		1
9		NEAR	NEAR	NO NEAR	NO NEAR		LOW	SHIFTRIGHT	1
10		NEAR	NEAR	NEAR	NEAR		ZERO	SHIFTRIGHT	1
11		NO NEAR	NEAR	NO NEAR	NO NEAR	FORWARD	HIGH	NOTURN	1
12	RIGHT	NEAR		NO NEAR		FORWARD	LOW	TURNRIGHT	1
13	LEFT	NEAR			NO NEAR	FORWARD	LOW	TURNLEFT	1
	connection type "AND"						connection type "AND"		

4 Motion Control

The control architecture of the wheeled mobile robot considered in the paper represents a modular hierarchy distributed structure. The proposed control system has two hierarchy levels, high and low. The high control level consists of a cognitive block (a knowledge-based block based on the fuzzy inference system presented in Fig. 5) coupled with a complementary model-based module. Such a controller takes into account the dynamics of the entire robotic system (2)-(20), including robot rigid-body dynamics, tire non-linear dynamics and tire-ground interaction effects (slipping, rolling resistance, etc.). The high control block is charged for sensor data acquisition, signal processing, sensor data fusion, sensor-based navigation, motion planning and the control of robot dynamics as well the appropriate control load distribution per particular robot wheels.

The low control block ensures the servo-control of DC-motors, whose task is the regulation of tire load (traction or braking torques) and tire angular velocities. In order to build the control for the wheeled mobile robot (Fig. 1) considered in the paper, the following assumptions have to be satisfied: (i) The model presented by relations (2 - 20) describes the system's physical behavior with satisfactory accuracy. (ii) The parameters of the model are acquired directly from the system by measurement or by estimation using corresponding sensor-based acquired information. (iii) In every sampling time it is possible to determine in a precise way the location (position) of the robot in the work-space as well as the location of the target point (Fig. 4) using corresponding localization sensors. (iv) Corresponding tire-ground interaction parameters (slipping and rolling resistance coefficients) can be estimated by use of the model described in Section 2. (v) Proximity sensors detect obstacles in the range of two meters with satisfactory accuracy.

If the previously listed assumptions are satisfied, the control system proposed in the paper is capable of ensuring high dynamic performance as well precise tracking of the target direction (Fig. 4). Concerning the accuracy of the model presented in Section 2, the relations (2 - 20) describe the entire system behavior (physics) including its empiric tire non-liner model. The tires of wheeled robots are factors that cause potential uncertainties in the system. Due to these reasons, their influence upon the system behavior is significant, and it is very important for the control efficiency that the tires' actual dynamics be modeled in an appropriate way. The parameters of rigid body model can be identified directly via measurement of the system or calculated based on appropriate measurements (e.g. robot mass, moments of inertia, dimensions, etc.). Tire-ground interaction parameters (adhesion parameters) are also estimated using appropriate experimental measurements and relations describing tire non-linear model.

The control algorithm of the wheeled mobile robot has to provide accurate path tracking and high dynamic performances of entire system. A control algorithm capable of providing such performance can be written in the following form [11]:

$$\begin{aligned} &\mathbf{T} = \mathbf{H}(\mathbf{q}) \cdot \hat{\ddot{\mathbf{q}}} + \mathbf{h}_{\mathbf{ccg}}(\mathbf{q}, \dot{\mathbf{q}}) - \mathbf{F}_{\mathbf{w}}(\dot{\mathbf{q}}), \\ &\hat{\ddot{\mathbf{q}}} = \ddot{\mathbf{q}}_{\mathbf{0}} - \mathbf{K}_{\mathbf{d}} \cdot (\dot{\mathbf{q}} - \dot{\mathbf{q}}_{\mathbf{0}}) - \mathbf{K}_{\mathbf{p}}(\mathbf{q} - \mathbf{q}_{\mathbf{0}}) \end{aligned} \tag{21}$$

where $\mathbf{K}_{\mathbf{p}}$ and $\mathbf{K}_{\mathbf{d}}$ are corresponding matrices of the proportional and differential control gains, while the other values appearing in (21) have been already explained in (8). From (11) the generalized forces T_X , T_Y and turning (spinning) torque T_ε are determined with respect to the absolute coordinate system (Fig. 1). Then, corresponding driving forces and torques in the longitudinal τ_x, lateral τ_y and yaw τ_ε direction are calculated from the relations:

$$\begin{aligned} \tau_x &= T_X \cos(\varepsilon) + T_Y \sin(\varepsilon), \\ \tau_y &= -T_X \sin(\varepsilon) + T_Y \cos(\varepsilon), \\ \tau_\varepsilon &= T_\varepsilon \end{aligned} \tag{22}$$

Generalized forces τ_x, τ_y and torque τ_ε are produced by corresponding tire forces F_x and F_y acting in the longitudinal and side directions (Fig. 2). The 2WD rover must be considered as an "over controlled" system since in the considered particular case there are four tire forces (F_{x1}, F_{y1}, F_{x2} and F_{y2}, shown in Fig. 2) while the global robot motion is performed in three particular coordinate directions: x, y and ε . The unknown forces $F_{xi}, i = 1,2$ and $F_{yi}, i = 1,2$ can be calculated indirectly (in a reverse way) from the relations (9) and (10), taking into account pre-determined generalized forces/torque τ_x, τ_y and τ_ε from (22). In order to perform one such procedure some additional relation must be provided to enable the calculation (by elimination of one unknown variable) of unknown tire forces F_{x1}, F_{y1} ,F_{x2} and F_{y2}. This auxiliary relation that enables a decrease in number of unknown variables is:

$$F_{y2} = \kappa_2 \cdot F_{y1} \tag{23}$$

where $\kappa_2 = F_{y2} / F_{y1}$ is the corresponding ratio (coefficient) that defines the relationship between the particular side force amplitudes F_{y1} and F_{y2} of the right and left side wheels. Taking into account relation (23), as well including it into (9) and (10), the corresponding three equations upon the three unknown variables F_{x1}, F_{x2} and F_{y1} can be solved from:

$$\begin{aligned} \tau_x &= F_{x1} + F_{x2}, \\ \tau_y &= (1 + \kappa_2) \cdot F_{y1} + F_{y3} + F_{y4}, \\ \tau_\varepsilon &= b \cdot F_{x1} - b \cdot F_{x2} + (1 + \kappa_2) \cdot l_x \cdot F_{y1} + F_{y3} \cdot l_f - F_{y4} \cdot l_r \end{aligned} \tag{24}$$

From (24) and auxiliary term (23) the unknown tire forces $F_{xi}, i = 1,2$ and $F_{yi}, i = 1,2$ are calculated. The tire forces $F_{yi}, i = 3,4$ appearing on the auxiliary tires are considered as disturbance of the system. The actual control variables of the wheeled mobile robot are not tire forces but particular angular velocities of the active wheels $\omega_i, i = 1,2$ (Fig. 1). Since the longitudinal $F_{xi}, i = 1,2$ and lateral tire forces $F_{yi}, i = 1,2$ are non-linear functions of the arguments, such as tire slip ratio s_i and tire slip angle α_i, this implies the necessity of an inverse procedure to be conducted in order to determine the actual control variables $\omega_i, i = 1,2$. Accordingly, the corresponding tire slip ratios $s_i, i = 1,2$ and tire slip angles $\alpha_i, i = 1,2$ should be calculated before.

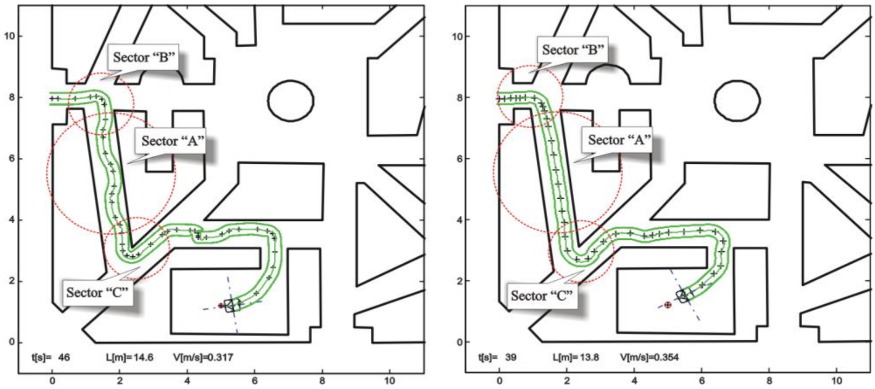
Theoretical aspects of modeling, spatial navigation, motion planning and control of robot dynamics considered in Sections 2, 3 and 4 will be verified in Section 5.

5 Verification of the Proposed Methodology

In order to verify the proposed navigation and control methodology, several characteristic simulation experiments are performed and analyzed in this section. An industrial, middle size indoor mobile robot platform RobuLab-10 (Fig. 1) is assumed for the purpose of simulation experiments whose model parameters are taken from the corresponding product-sheet [12] and given in the Appendix of the paper. The robot chosen is simulated moving in conditions of unknown environment, modeled as kind of an indoor labyrinth scenario (Fig. 7), and in conditions of confined work-space with the appearance of different ambient contingency risks, such as: variable tire-ground interaction characteristics (e.g. variable slipping coefficient and rolling resistance) and irregular geometry (e.g. obstacle shapes the and sizes of surrounding obstacles). For this purpose, the Virtual WRSN, a specialized modeling/simulation software toolbox for Matlab/Simulink, was used [14].

The first considered simulation example brings together and compares the characteristics of two concurrent control approaches to be applied with mobile wheel-based robots, kinematic (differential steering) and a dynamic (skid steering) approach. Differential steering is based on the calculation of referent right and left tire angular velocities and directly from (2), neglecting the effects of slipping of the robot tires and neglecting their distinct non-linear nature. The concurrent skid steering approach takes into account the entire rigid-body robot dynamics, non-linear tire dynamics and variable tire-ground interaction characteristics such as tire slipping and rolling resistance.

In order to provide an appropriate comparison of two considered concurrent control approaches, the simulation conditions assumed should be equal (same) in both test cases. This means the same navigation strategy as well as the same motion conditions (tire adhesion characteristics) are considered during the simulation tests. The slipping coefficient in the considered simulation examples is assumed to be $\mu = 0.75$. This corresponds to a dry, relatively high-adhesive tire-ground surface.



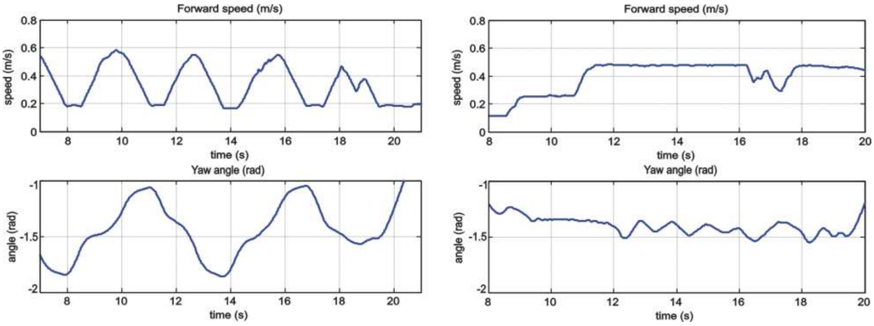
a) b)

Figure 7

Simulation of wheel traces of the mobile RobuLab-10 robot (Fig. 1) obtained in the cases:
a) path realized by control algorithm based on use of the kinematical approach, and
b) path obtained by means of implementation of integrated (dynamic) control

The curvilinear motion of the mobile robot RobuLab-10 in an assumed labyrinth scenario (Fig. 7) was performed from the start-point towards the target-point along the target direction in the presence of surrounding obstacles of different shape and size. Under the same conditions, the robot-rover is controlled by use of dynamic control approach, too. The corresponding wheel traces of one such motion is shown in Figs. 7a and 7b. By comparison of performances of the two chosen test motions shown in Figs. 7a and 7b, better (smoother) motion appears in the case when a dynamic control is applied than in the case when a simplified kinematical algorithm is used. Consequently, robot with kinematicaly-based control makes the trip in approximately 46 (s) while the same robot needs less time (approximately 39 (s)) with the use of the dynamic control algorithm to perform the same task. The consequence is that the dynamically controlled robot moves noticeable faster than the same robot controlled in a kinematically-based way. The robot behaviour corresponding to the cases presented in Figs. 7a and 7b, is shown in Figs. 8a and 8b by presentation of the corresponding state variables as characteristic system state indices.

With the analysis of these two motions, the following conclusion can be brought out. The dynamic control of robot motion ensures comparatively better system performance and maneuverability than the concurrent kinematically-based algorithm. It is characterized by more smooth and faster motion along the path. In addition, in cases of cornering manoeuvres (shown in Sectors B and C, Figs. 7a and 7b), the robot keeps (tracks) an imagined middle line of the corridor in a better way when it is controlled by the dynamic-based algorithm proposed in the paper.



a) b)

Figure 8

Robot state variables captured during the motion through the Sector "A" (Figs. 7): a) actual forward speed and yaw-rate of robot obtained in the case of kinematical approach applied to control robot motion, b) the same state indices obtained for the case of dynamic approach to control the same system

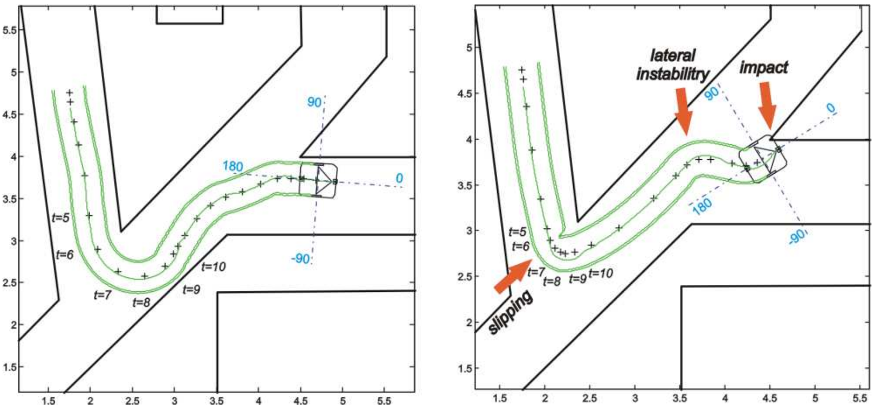
In such a way, by use of dynamic approach to control robot motion, less risk of collision with surrounding obstacles is attained, especially in the cornering manoeuvres. The reason why the dynamic control approach provides better system performance than the concurrent kinematical approach can be explained in the following way. By calculation of the referent tire angular velocities and directly from the robot kinematic model (2), the tire non-liner effects as well as the influence of tire-ground interaction factors are not taken in account.

The second simulation example considered in the paper regards the case when tire adhesion parameters vary during robot motion. Instead of the previously considered $\mu = 0.75$, a new decreased friction coefficient of $\mu = 0.45$ is introduced as a potential contingency risk of the robot motion. It corresponds to the case of a very slippery, low-friction floor surface. The corresponding simulation results that demonstrate one such dynamic behaviour are presented in Figs. 9a and 9b. Complementary simulation results presented in Figs. 9a and 9b are given to support the aforementioned theoretical considerations as well as to highlight the benefits of implementation of the proposed dynamic control with respect to the simplified kinematical approach to control motion.

Tire angular velocities are calculated by taking into account the linear tire speed, tire slip ratio and tire slip angle. The obtained right and left tire angular velocities $\omega_i, i = 1,2$ for robot motion presented in Fig. 9 are shown in Fig. 10.

Conclusions

The paper regards sensor-based, non-visual navigation and integrated control of indoor ambient adaptive wheel-based robots in unknown environments with contingency risks. The proposed architecture couples at the high functional level two specialized modules: a knowledge-based block and a model-based block. The cognitive knowledge-based block is synthesized for sensor-based navigation and spatial reasoning while the complementary model-based module is dedicated to the integrated control of robot motion and system dynamics.

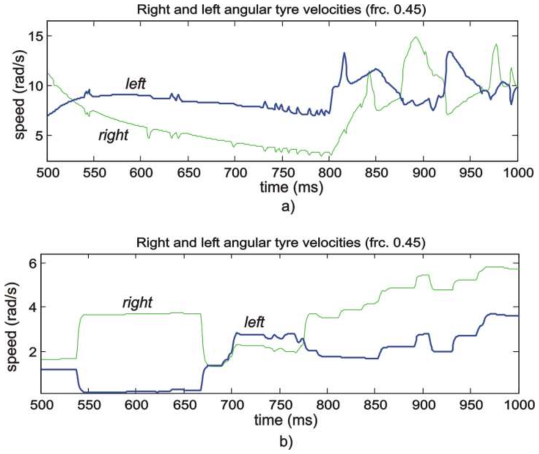


a)

b)

Figure 9

Wheel traces obtained in simulation tests for a characteristic "S"cornering maneuver (Sector C, Figs. 7a and 7b) in a case of low friction conditions (0.45): a) case when the system is controlled by implementation of dynamic, and b) simplified kinematical approach



a)

b)

Figure 10

Comparison of the tire angular velocities ω_i, $i = 1,2$ observed for a fragment of robot motion (Fig. 9) on the low-friction ground surface ($\mu = 0.45$): a) dynamic, and b) kinematical control approach

Navigation and control modules are coupled within a unique robot controller that is designed to ensure non-visual, target-oriented accurate navigation in unknown environments with contingency risks. The risks regard to variable tire-road interaction conditions, such as uncertainty in slipping conditions and rolling resistance as well as in variety of obstacle shapes and sizes. The cognitive navigation module must enable reliable, collision free motion in the presence of different obstacles. The proposed integrated control of a mobile robot is designed to improve dynamic performances (longitudinal and lateral stability, maneuverability in a confined and cluttered environment, etc.) of the robotic system and to ensure the accurate tracking of target direction towards a goal point. The paper considers the non-linear form of tire model and tire-ground interaction effects and proposes how this model can be effectively used for estimation of contingency risks and compensation of their influence upon the system. The aforementioned control structure ensures better system robustness to the system uncertainties of different type and better implementation capabilities of indoor mobile robots.

Acknowledgement

The results presented in the paper are part of the research project ‘Research and Development of Ambientally Intelligent Service Robots of Anthropomorphic Characteristics’, TR-35003 funded by Ministry of Science of Republic Serbia and partially funded by the Provincial Secretariat for Science and Technological Development of Autonomous Province of Vojvodina, Republic of Serbia, under the contract no. 114-451-2116/2011. Project is also supported by the SNSF SCOPES IP project ‘CARE-robotics’, ID: IZ74Z0_137361/1.

References

[1] Morin, P., Samson, C.: Motion Control of Wheeled Mobile Robots. In: Siciliano, B. and Khatib, O. (eds.) Handbook of Robotics, Springer, 2008, pp. 729-825

[2] Minguez, J., Lamiraux, F., Laumond, J-P.: Motion Planning and Obstacle Avoidance. In: Siciliano, B. and Khatib, O. (eds) Handbook of Robotics Springer, 2008, pp. 826-850

[3] Lumelsky, V., Stepanov, A.: Path Planning Strategies for a Point Mobile Automation Moving Admist Unknown Obstacles of Arbitrary Shape. Algorithmica 2, 1987, pp. 403-430

[4] Khatib, O.: Real-Time Obstacle Avoidance for Manipulators and Mobile Robots. International Journal of Robotic Research 5, 1986, pp. 90-98

[5] Minguez, J., Montano, L.: Nearness Niagram (ND) Navigation: Collision Avoidance in Troublesome Scenarios. IEEE Transact. on Robot. Autom. 20(1), 2000, pp. 45-59

[6] Minguez, J., Osuna, J., Montano, L.: A Divide and Conquer Strategy to Achieve Reactive Collision Navigation Systems. IEEE Int. Conf. Robot. Autom., 2004, pp. 412-418

[7] István Nagy, Behaviour Study of a Multi-Agent Mobile Robot System during Potential Field Building, Acta Polytechnica Hungarica, Vol. 6, No. 4, 2009, pp. 111-136

[8] Radu-Emil Precup, Stefan Preitl, Emil M. Petriu, József K. Tar, Marius L. Tomescu, Claudiu Pozna, Generic Two-Degree-of-Freedom Linear and Fuzzy Controllers for Integral Processes, Journal of Franklin Institute-engineering and Applied Mathematics, Vol. 346, No. 10, 2009, pp. 980-1003

[9] Kovacs, S.; Vincze, D.; Gacsi, M.; Miklosi, A.; Korondi, P., Ethologically Inspired Robot Behavior Implementation, 4th International Conference on Human System Interactions (HSI), 2011, pp. 64-69

[10] Károly Nagy, Szabolcs Divéki, Péter Odry, Matija Sokola, Vladimir Vujičić, A Stochastic Approach to Fuzzy Control, Acta Polytechnica Hungarica, Vol. 9, No. 6, 2012, pp. 29-48

[11] Aleksandar Rodić, Miomir Vukobratović: Dynamics, Integrated Control and Stability of Automated Road Vehicles. Ibidem Verlag, Stuttgart, 2002

[12] RobuLAB10 product sheet. Robosoft. http://www.robosoft.fr/img/data/robuLAB10_web.pdf, Accessed May 2011

[13] Pacejka, H., Bakker, B.: The Magic Formula Tire Model. Vehicle System Dynamics, Vol. 21, 1993

[14] Aleksandar Rodic, Gyula Mester, "Virtual WRSN – Modeling and Simulation of Wireless Robot-Sensor Networked Systems". Proceedings of the 8th IEEE International Symposium on Intelligent Systems and Informatics, SISY 2010, Subotica, Serbia, 2010, pp. 115-120

[15] Experimental testbed station for research and development of robot-sensor networked systems, Robotics Laboratory, Institute Mihailo Pupin, http://www.pupin.rs/RnDProfile/robots.html. Accesed January, 2013

[16] Gyula Mester, Intelligent Mobile Robot Motion Control in Unstructured Environments, Acta Polytechnica Hungarica, Journal of Applied Sciences, Vol. 7, Issue No. 4, Budapest, Hungary, 2010, pp. 153-165

[17] Aleksandar Rodić, Khalid Addi, Mirko Jezdimirović, Sensor-based Intelligent Navigation and Control of Autonomous Mobile Robots in Advanced Terrain Missions, Scientific Technical Review, Vol. 60, No. 2, 2010, pp. 7-15

Appendix

Kinematical and dynamic parameters of the industrial mobile robot platform RobuLab-10 [8] was assumed for modeling and simulation in this paper. Parameters are taken from the product sheet or estimated by use of the model. Control gains are synthesized in a manner described at the end of Appendix. Parameters used in the paper are presented in Table A1. Tire model coefficients $B_j, C_j, D_j, E_j,\ j = 1,2,3$ whose values are given in Tab. A1, are determined from the corresponding empiric tire force graphs. Index $j=1$ is used for the longitudinal tire force model, $j=2$ for the lateral tire model and $j=3$ for the tire self-aligning torque model curves.

Control gains of the PD regulator to be used in (21) are determined by implementing of the Pole Placement Method.

Table A1

Kinematical, dynamic and control parameters used in simulation

Parameter	Value	Unit
r_t	0.05	m
l_f	0.215	m
b	0.360	m
l_r	0.235	m
l_x	0	m
m	20	kg
I_z	0.6042	$kg\,m^2$
g	9.81	m/s^2
K_x	1.36	$N\,s^2/m^2$
K_y	1.50	$N\,s^2/m^2$
f_{ri}	0.02	
μ_s	0.75	
B_1, B_2, B_3	0.0985 , 0.1884 , 0.4966	
C_1, C_2, C_3	1.65 , 1.30 , 2.40	
D_1, D_2, D_3	38.84 , 38.84 , 12.00	N
E_1, E_2, E_3	(-7.4617 , -2.8556 , -0.1006) x $10^5 \times 10^5$	
k_p	355.3058	Nm/rad
k_d	37.6991	$Nm\,s/rad$

The frequency and the relative damping coefficient of the close-loop control system are chosen to be $f_{PD}=3.00 \quad (Hz)$ and $\zeta_{PD}=1.00$. That ensures the poles of the control system are in the left plane. The following relations are used to determine the PD control gains (obtained values are given in Table A1):

$$\varpi = 2\pi \cdot f_{PD}$$

$$k_p = \varpi^2, \quad k_d = 2 \cdot \zeta_{PD} \cdot \varpi$$

$$K_p = \operatorname{diag}\{k_p\}, \quad K_d = \operatorname{diag}\{k_d\}$$

Co-Exceedances in Eurozone Sovereign Bond Markets: Was There a Contagion during the Global Financial Crisis and the Eurozone Debt Crisis?

Silvo Dajčman

University of Maribor, Faculty of Economics and Business, Razlagova 14, 2000 Maribor, Slovenia, e-mail silvo.dajcman@uni-mb.si

Abstract: The paper examines contagion between the sovereign bond markets of six Eurozone countries (France, Germany, Ireland, Italy, Spain, and Portugal) in the period from January 2000 to August 2011. A multinomial logistic model is applied to analyze contagion based on measuring joint occurrences of large yield changes (i.e., co-exceedances), while controlling for developments in common and regional factors that affect all sovereign bond markets simultaneously. I found that the Eurozone's stock markets (EUROSTOXX50) returns, United States' Treasury note yields, and the Euro-U.S. dollar (EUR-USD) exchange rate significantly impact the probability of extreme positive yield moves in the Eurozone's sovereign bond markets. Positive EUROSTOXX50 returns and upside moves in U.S. Treasury note yields increased the probability of extreme positive sovereign bond yield moves in the Eurozone, whereas an increase in the EUR-USD exchange rate significantly reduced the probability. Conditional volatility in the Eurozone stock markets and the money market interest rate do not significantly impact the probability of extreme yield increases in the Eurozone's sovereign bond markets. Furthermore, the probability of observing exceedance across Eurozone sovereign bond markets increased dramatically during the Eurozone debt crisis compared to the pre-crisis period. This study's results also indicate less synchronous extreme yield dynamics across the Eurozone sovereign bond markets during the global financial crisis, especially during the Eurozone debt crisis compared to the pre-crisis period.

Keywords: Sovereign bond markets; Eurozone debt crisis; Contagion

1 Introduction

In recent years, European countries have been hit by two episodes of major financial market distress: the global financial crisis and the sovereign debt crisis. The Eurozone sovereign debt crisis, triggered by mounting concerns about the fiscal sustainability of Mediterranean countries, led to a further surge in sovereign bond yields. The shock spilled over to other Eurozone sovereign debt markets, thereby raising the question whether public debts across the Eurozone are sustainable. Prompted by financial market pressures, large-scale fiscal austerity measures have been announced in practically all Eurozone countries and sovereign debt management has advanced to the top of the international policy agenda.

As [16] argued, quantifying the exposure of developed countries to sovereign bond market spillovers (i.e., exposure to contagion) can help policymakers gain insight into overall financing constraints, as well as the external risks an economy faces. By analyzing contagion, knowledge is gained regarding whether a shock in one segment of a national financial market is transmitted across markets via channels that appear only during turbulent periods, or whether these shocks are transmitted via channels or inter-linkages that exist in all states of the world (non-crisis or crisis periods). [11] noted that the effectiveness of economic policy measures aimed at reducing a market's vulnerability to contagion will depend on whether the contagion occurred as a result of the transmission of shocks through pre-existing, long-term links or through crisis contingent channels.

The literature includes many definitions of contagion (see e.g. [2], [4], [5], [9], [18]). [9] provides one of the most commonly accepted definitions of contagion, namely the "shift contagion", which regards contagion as a shift or change in how shocks spread from one country (or asset class) to another during normal periods (pre-crisis) and how during crisis periods.[1] A common way to measure contagion is through the conditional correlation changes between the returns of asset classes. Using this approach, contagion is identified if conditional correlation significantly increases in the crisis period in relation to the tranquil (non-crisis) periods. This method has been applied both empirically and extensively (e.g. by [4], [10]).

Correlations that give equal weight to small and large returns, however, are not appropriate to evaluate the differential impact of large returns (or yields in the case of bonds). As [1] argued, when large shocks exceed some threshold they can generate panic and propagate across countries. This propagation, however, is hidden in correlation measures by the large number of days when little of importance happens. Furthermore, the correlation coefficient is not an adequate measure of co-movement or interdependence and is difficult to interpret due to its sensitivity to heteroskedasticity (see [14] and [10]). The correlation coefficient is also a linear measure that is inappropriate if contagion is not a linear phenomenon but rather an event characterized by nonlinear changes in market associations [1].

[1] Contagion must be distinguished from interdependence. As [10] argued, if two markets are traditionally highly correlated, and the correlation does not increase significantly after a shock in one market, then any continued high level of market co-movement suggests strong and real linkages between the two economies. In this case, there is no contagion but only high interdependence.

Rather than computing correlations of bond yield changes, here I base the analysis of contagion in sovereign bond markets on a measure of the joint occurrences of large positive yield changes (i.e., co-exceedances), an extreme value theory concept that [1] introduced. Exceedance is defined as an occurrence of a large bond yield change, that is, one above a certain threshold. Co-exceedances, on the other hand, are joint exceedances of two financial market returns above a certain threshold. This measure circumvents problems associated with the correlation coefficient because co-exceedances are not biased in periods of high volatility and are not restricted to modeling linear phenomena (see [3] and [6]).

A multinomial logistic regression can be used to model the occurrences of large bond yield changes. An important advantage of multinomial logistic analysis is that one can condition on attributes and characteristics of the exceedance events using control variables (or covariates) measured with information available up to the previous day. Following [1], the strength of contagion between sovereign bond markets is then measured as the fraction of co-exceedance of extreme positive bond yield changes that are not explained by the covariates included in the model.

In the present paper, I use a method developed by [1] to measure the strength of contagion between the sovereign bond markets of six Eurozone countries (France, Germany, Ireland, Italy, Spain, and Portugal) in the period from January 2000 to August 2011. A multinomial logistic model is applied to measure contagion between sovereign bond markets in a pair-wise manner. In other words, contagion is measured between pair-wise observed sovereign bond markets. To separate contagion from interdependence, I include more covariates in the multinomial logistic model than did [1] following suggestions in the empirical literature on contagion in the financial markets ([6], [7]). This includes the average stock market returns of the Eurozone (proxied by the returns on the EUROSTOXX50 index); the conditional volatility of the EUROSTOXX50 returns modeled as EGARCH(1,1); Eurozone money market interest rate level (3-month EURIBOR); U.S. Treasury note yield changes; and returns on the Euro-U.S. dollar (EUR-USD) exchange rate. The response of probability estimates to the full range of values associated with different covariates are also computed and presented graphically to inspect whether the relationship between the probability of (co-)exceedances and covariates are linear or nonlinear. A multinomial logit model is specified in a way that enables us to investigate whether the most recent episodes of financial market distress (i.e., the global financial crisis and the Eurozone debt crisis) significantly impacted the probability of contagion in the investigated Eurozone sovereign bond markets.

2 Methodology

Exceedances in terms of extreme positive sovereign yield changes in a particular country and pair-wise joint occurrence (i.e., joint occurrence in two observed sovereign bond markets or co-exceedance) of extreme positive sovereign bond yield changes can be modeled as a polytomous variable. The dependent polytomous variable at time t (y_t; $t = 1, \dots, T$) in the present paper can fall into one of three categories ($j = 1,2,3$): no exceedance in any of the pair-wise countries ($j = 1$); exceedance observed in one of the countries in the pair ($j = 2$); and co-exceedance. This third category represents a simultaneous exceedance in both the countries, representing contagion ($j = 3$). Probabilities associated with the events captured in the polytomous variables can then be estimated using a multinomial logistic model ([1]). An advantage of multinomial logistic analysis is that one can condition on attributes and characteristics of the exceedance events using control variables (explanatory variables or covariates) that are measured using information available up to the previous day.

The multinomial logit model assumes that the probability of observing category j (of the three possible categories) in the dependent polytomous variable, P_j, is given by Equation (1) ([11])

$$P_j = \Pr(y_t = j) = \frac{\exp(\beta_j' \boldsymbol{x})}{1+\sum_{k=2}^{3} \exp(\beta_k' \boldsymbol{x})}, \tag{1}$$

where $\boldsymbol{x}$ is a $T \times n$ matrix of covariates (with n being the number of different covariates) and β the vector of coefficients (including a constant) of a particular category associated with the covariates.[2] The covariates included in the model are the average stock market returns of the Eurozone (proxied by returns on the EUROSTOXX50 index); the conditional volatility of the EUROSTOXX50 returns modeled as EGARCH(1,1)[3]; the Eurozone money market interest rate level (3-month EURIBOR); 10-year U.S. Treasury note yield changes; and returns on the EUR-USD exchange rate. Because I also want to answer the question of whether the probability of contagion increases in a crisis period compared to a non-crisis period, also two dummy variables are included.[4] The first dummy variable

[2] To separate contagion from interdependence, it is important to identify common and regional factors that impact all countries simultaneously ([6]). A failure to model common and regional factors may result in tests of contagion being biased toward a positive finding of contagion.

[3] The EGARCH model of [16] stipulates that negative and positive returns have different impacts on volatility.

[4] Changes in Treasury note yields and the EUR-USD exchange rate (log) returns are included as a proxy for global macroeconomic developments and the associated inflation, liquidity, and credit risks (see e.g. [10] and [6]; [16]). The region-specific factors that capture local financial market conditions are the Eurozone money market rate, EUROSTOXX50 index, and its conditional volatility. As argued by [7], the bond markets should not be studied in isolation, because there are interaction effects across

represents the crisis period from September 16, 2008[5] to April 22, 2010 and the second represents the crisis period from April 23, 2010 to August 31, 2011[6].

Coefficients β are specific to each category, so that there are $j \times n$ coefficients to be estimated. The coefficients are not all identified unless one imposes a normalization (see [12]). Normalization in the present paper is achieved by setting the coefficient of the first category ($j = 1$) to be zero. All regression coefficients of Equation (1) are thus calculated with respect to the first category (category 1) as a base category.

The model is estimated using maximum likelihood with the log-likelihood function for a sample of t observations given by

$$lnL = \sum_{t=1}^{T} \sum_{j=1}^{3} d_{tj} \log(P_{tj}), \tag{2}$$

where d_{tj} is a dummy variable that takes a value one if observation t takes the jth category and zero otherwise. Because P_{tj} is a nonlinear function of the βs, an iterative Newton-Rahpson's estimation procedure is applied. Goodness-of-fit is measured using the pseudo-R^2 of [15] where both unrestricted (full model) likelihood, L_ω, and restricted (constants only) likelihood, L_Ω, functions are compared

$$pseudo\ R^2 = 1 - (\frac{logL_\omega}{logL_\Omega}). \tag{3}$$

After calculating regression coefficients, the probabilities of each of the three categories, Pj, are computed by evaluating the covariates at their unconditional values

$$P_j = \frac{\exp(\beta_j' x^*)}{1 + \sum_{k=2}^{3} \exp(\beta_k' x^*)}, \tag{4}$$

where x^* is the vector of the unconditional mean values of the covariates. Because the coefficients in a multinomial logit model are difficult to interpret, following [12] and [1], the marginal changes in probability for a given unit change in the independent covariate (i.e., marginal effects) are calculated and tested whether they are significantly different from zero. The marginal effects (δ_j) are given by the following equation (see [12]):

$$\delta_j = \frac{\partial P_j}{\partial x}\Big|_{x=x^*} = P_j \left[\beta_j - \sum_{k=1}^{3} P_k \beta_k\right]\Big|_{x=x^*}. \tag{5}$$

different asset classes. In their study, [1] included only conditional volatility of the stock market, exchange rate returns, and the interest rate level.

[5] On September 16, 2008 the investment bank Lehman Brothers collapsed and started the global financial crisis.

[6] On April 23, the Greek government requested a bailout from the EU/IMF. I take this date as the start of the sovereign debt crisis in the Eurozone. August 31, 2011 is the end of the observation period in the paper.

[1] noted that it is often difficult to judge whether changes in probabilities of a given category are economically large or small. In the present paper, therefore, I also present the responses of probability estimates to the full range of values associated with different covariates, rather than just at its unconditional means, and present them graphically.

3 Data and Empirical Results

Extreme upper tail yield behavior in the sovereign bond markets in the six Eurozone countries, listed in Table 1, is analyzed based on the sovereign bond yield changes. The daily changes of bond yields were calculated from the yields (y) of central-government bonds (bullet issues) with 10 years maturity as $ln(y_t) - ln(y_{t-1})$.[7] Days with no trading in any of the observed market were left out. Yield changes (and all other variables, i.e. covariates) are calculated as two-day rolling-average logarithmic changes in order to control for the fact of the different open hours of the markets on which the variables in the model are formed.[8] The data for bond yields are from the Denmark's central bank.[9] Table 1 presents some descriptive statistics of the data.

Table 1

Descriptive statistics of bond yield changes

	Period of observation	Min	Max	Mean	Std. deviation	Skewness	Kurtosis	Jarque-Bera statistics
France	3 January 2000 – 31 August 2011	-0.0492	0.0600	-0.000220	0.01059	0.1360	4.7921	407.3***
Ireland	3 January 2000 – 31 August 2011	-0.215	0.0846	0.000139	0.01237	-1.3056	38.1730	15,419.9***
Italy	3 January 2000 – 31 August 2011	-0.1406	0.0753	-0.00004	0.009924	-0.6834	19.7355	34,949.3***
Germany	3 January 2000 – 31 August 2011	-0.0760	0.0764	-0.000303	0.01208	0.0345	6.3872	1,422.8***
Portugal	3 January 2000 – 31 August 2011	-0.3006	0.1449	0.000226	0.01358	-3.3664	93.2459	1.015,175.3 ***
Spain	3 January 2000 – 31 August 2011	-0.1582	0.0607	-0.000039	0.01101	-1.2329	23.6001	53,357.3***

Notes: Jarque-Bera statistics: *** indicate that the null hypothesis (of normal distribution is rejected at a 1% significance level, ** that null hypothesis is rejected at a 5% significance level and * that the null hypothesis is rejected at a 10% significance level.

[7] Bond yield changes are calculated the same way as in [8] and [13].

[8] The same approach is used by [10].

[9] The data series for the 3-month EURIBOR and the EUR-USD dollar exchange rate were obtained from the web page of Deutsche Bundesbank. The data series of EUROSTOXX50 and the 10-year U.S. Treasury note yields are from Yahoo! Finance.

All series display significant leptokurtic behavior as evidenced by large kurtosis with respect to the Gaussian distribution. The Jarque-Bera test rejects the hypothesis of a normally distributed observed time series.[10]

Table 2 reports Pearson's correlation coefficients of the two-day rolling-average logarithmic bond yield changes. The greatest linear co-movement[11] of bond yield changes in the observed period was achieved between the French-German and between the Italian-Spanish sovereign bonds, while between the German-Portuguese and German-Irish sovereign bond yields the smallest correlation is observed.

Table 2

Pearson's correlation of sovereign bond yield changes between Eurozone countries

	France	Germany	Ireland	Italy	Portugal	Spain
France	1					
Germany	0.9214	1				
Ireland	0.5277	0.3906	1			
Italy	0.6856	0.5334	0.7063	1		
Portugal	0.4690	0.3288	0.8089	0.6854	1	
Spain	0.6641	0.5286	0.7433	0.9048	0.7299	1

Notes: all the correlation coefficients are significantly different from zero.

Following [1] an extreme positive yield change or exceedance is defined as the one that lies above the 95th quintile of the marginal yield change distribution. In Table 3, the count numbers of exceedances and joint occurrences of extreme returns (co-exceedances) are reported. The results in Table 3 are presented as a lower triangular matrix, with the diagonal entries representing the number of exceedances in the particular country (because only the upper 5% of the extreme bond yield changes are of interest to the present study, there are $0.05 * 2974 = \|148{,}7\| = 149$ exceedances). The other fields in the lower triangular matrix, presented in Table 3, are the counts of pair-wise co-exceedances of daily bond yield changes.

Table 3

Statistics of the counts of the (co-)exceedances of daily bond yield changes

	France	Germany	Ireland	Italy	Portugal	Spain
France	149					
Germany	118	149				
Ireland	71	68	149			
Italy	97	80	77	149		
Portugal	64	61	97	75	149	
Spain	96	86	89	111	83	149

Notes: A total of 2,974 daily observations of two-day rolling window bond yield changes are observed. The numbers in the table can be explained as follows. For France, there were 149 occurrences (days) when the two-day rolling-window yield changes exceeded the 95th quintile of the marginal yield change distribution in the total observed period. Further, there

[10] The stationarity of bond yield changes was also examined, but the results (they lead to rejection of the unit root) are not relevant for this study, as I am interested only in the upper 5% of the bond yield changes distribution, and are thus not reported.

[11] The Pearson's correlation is a linear measure of comovement.

were 118 days of joint exceedances (i.e., co-exceedances) in the sovereign bond markets of France-Germany, 71 days of co-exceedances in France-Ireland sovereign bond markets, etc.

The greatest count of co-exceedances is achieved for the following pairs of national sovereign bond markets: Germany-France, Italy-Spain, France-Italy, and Ireland-Portugal. The lowest counts of sovereign bond markets are achieved for the pairs of France-Portugal and Germany-Portugal. Figure 1 in the Appendix illustrates the time series of (co-)exceedances for these pair-wise observed sovereign bond markets.[12]

Notably, in the period from mid-2000 to mid-2001 and in the year 2007, there were almost no exceedances or co-exceedances for the countries investigated and illustrated in Figure 1. After the global financial crisis began, in the third quarter of 2008, the count of (co-)exceedances across all observed pairs of sovereign bond markets increased. It is also evident that after the start of year 2010, the count of outcome 2 (exceedance) increased for the sovereign bond markets of France–Portugal and Germany–Portugal and dominated over outcome 3 (co-exceedance). This clearly indicates that extreme (positive) yield dynamics was achieved in only one of the countries in the investigated pair, namely Portugal. Judging just from Figure 1 and not controlling for the effects of the control variables, contagion in the sovereign bond markets would be identified when the counts of outcome 3 increased compared to non-crisis periods. Between the sovereign bond markets of France and Germany, contagion would then be identified in years 2003, 2009, and 2011 and between the sovereign bond markets of France and Italy in 2003 and 2009. Similarly, episodes of contagion in other sovereign bond markets could also be identified.

As argued in the Introduction and Section 2 of the present paper, to separate contagion from interdependence, one must control for the effects of the control variables. In the present paper, this is achieved by estimating multinomial logistic model (1). Results of the model are reported in Tables 4a and 4b.

Table 4a

Estimates of the multinomial logit regression model (1) for specific pair-wise observed sovereign bond markets

	Fra-Ger	Fra-Ire	Fra-Ita	Fra-Por	Fra-Spa	Ger-Ire	Ger-Ita	Ger-Por
Outcome 2								
Constant	-4.881^a	-3.4170^a	-5.430^a	-3.419^a	-4.339^a	4.420^a	-4.792^a	-3.872^a
EUROSTOXX50 (returns)	53.650^a	13.341	7.945	16.080^c	19.766^c	25.304^a	16.066^c	25.499^a
cond. volatility of EUROSTOXX50 returns	148.895	359.471	325.581	262.860	619.915	175.443	186.174	238.669
EURIBOR (level)	-0.029	-0.233^c	0.182	-0.219^c	-0.224	-0.067	0.114	-0.188

[12] In total I investigate (co-)exceedances for 15 ($=\frac{6*5}{2}$) pairs of national sovereign bond markets.

U. S. 10y T.N. yield changes	90.870^{a}	56.763^{a}	59.529^{a}	60.727^{a}	51.617^{a}	53.234^{a}	74.889^{a}	62.105^{a}
EUR-USD returns	-23.302	-82.663^{a}	-29.827	-96.70^{a}	-118.543^{a}	-41.258^{b}	-19.897	-60.050^{a}
Crisis period 1	0.457	0.666^{c}	2.024^{a}	0.943^{a}	0.937^{b}	1.913^{a}	1.6234^{a}	1.625^{a}
Crisis period 2	1.707^{a}	2.373^{a}	3.402^{a}	2.366^{a}	2.911^{a}	3.474^{a}	3.050^{a}	2.984^{a}
Outcome 3								
Constant	-4.322^{a}	-4.608^{a}	-4.015^{a}	-4.646^{a}	-3.981^{a}	-4.503^{a}	-4.333^{a}	-4.732^{a}
EUROSTOXX50 (returns)	45.774^{a}	21.124^{b}	20.135^{b}	26.707*^{b}	26.419^{a}	20.514^{b}	25.878^{a}	28.807^{a}
cond. volatility of EUROSTOXX50 returns	137.755	850.377^{b}	898.133^{b}	754.001^{c}	597.595	788.804^{c}	909.479^{b}	612.837
EURIBOR (level)	-0.065	-0.040	-0.1148	-0.0703	-0.106	-0.0914	-0.1054	-0.0578
U. S. 10y T.N. yield changes	131.133^{a}	108.138^{a}	120.650^{a}	106.758^{a}	113.659^{a}	115.404^{a}	122.586^{a}	108.948^{a}
EUR-USD returns	3.979	-52.205^{b}	-69.067^{a}	46.594^{c}	-36.297	-47.287*	-53.392^{b}	-43.705
Crisis period 1	-0.064	-.0419	-1.146^{a}	-0.069	-0.579	-0.630	-0.749^{c}	0.028
Crisis period 2	0.726^{c}	0.358	-0.913^{c}	0.195	-0.730	0.023	-0.772	0.247
Log likelihood	-562.86	-715.22	-643.57	-718.59	-635.19	-669.04	-667.29	-681.90
LR chi (14)	464.67	455.87	462.49	476.30	485.63	560.34	510.97	560.14
Prob.>chi2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo-R^{2} (McFadden)	0.292	0.241	0.264	0.249	0.277	0.292	0.277	0.291

Notes: Estimates of the regression coefficients of model (1) are given. EUROSTOXX50 (returns) are the two-day, rolling-average log returns of the EUROSTOXX50, whereas cond. volatility of EUROSTOXX50 returns are the EGARCH (1,1) conditional volatilities of the EUROSTOXX50 two-day rolling average log returns. U.S. 10y T. N. yield (log) changes are the two-day, rolling-average log changes of the U.S. Treasury note yields. Crisis 1 is a time dummy for the first crisis period (September 16, 2008 – April, 22, 2010) and Crisis 2 is a time dummy of the second crisis period (from April 23, 2010 – August 31, 2011). Outcome 1 (no (co-)exceedance) is the base category. Outcome 2 presents the results of model (1) for category 2 (i.e., exceedance in one country only), whereas outcome 3 presents the results of model (1) for category 3 (i.e. co-exceedance). $^{a}/^{b}/^{c}$ denote the 1%, 5%, and 10% significance of the rejection of the null hypothesis that the regression coefficient is equal to 0, based on z-statistics. LR chi (14) reports the likelihood-ratio chi-square test (at 14 degrees of freedom) that for both equations (i.e., for outcome 2 and outcome 3) at least one of the covariate's coefficients is not equal to zero. Prob. > chi2 reports the probability of getting a LR test statistic as extreme as, or more so, than the observed under the null hypothesis (i.e., that all of the regression coefficients across both models [i.e. for outcome 2 and outcome 3] are simultaneously equal to zero).

Table 4b

Estimates of the multinomial logit regression model (1) for specific pair-wise observed sovereign bond markets

	Ger-Spa	Ire-Ita	Ire-Por	Ire-Spa	Ita-Por	Ita-Spa	Por-Spa
Outcome 2							
Constant	-4.777^{a}	-4.054^{a}	-4.109^{a}	-4.098^{a}	-3.148^{a}	-5.316^{a}	-3.682^{a}
EUROSTOXX50 (returns)	27.158^{a}	2.123	18.287^{c}	5.724	2.487	-0.794	5.080
cond. volatility of EUROSTOXX50 returns	297.095	455.702	281.783	82.863	277.361	150.641	260.602
EURIBOR (level)	-0.064	0.004	-0.174	-0.104	-0.254^{b}	0.163	-0.196
USA 10y T.N. yield changes	53.659^{a}	34.557^{a}	25.668^{a}	26.014^{a}	42.949^{a}	44.200^{a}	34.415^{a}
EUR-USD returns	-78.945^{a}	-54.392^{a}	-72.904^{a}	-80.411^{a}	-87.045^{a}	-23.325^{a}	-112.409^{a}
Crisis period 1	1.765^{a}	1.085^{a}	1.719^{a}	1.466^{a}	0.649^{c}	2.014^{a}	1.117^{a}
Crisis period 2	3.495^{a}	2.695^{a}	2.537^{a}	2.826^{a}	1.962^{a}	2.885^{a}	2.434^{a}

Outcome 3							
Constant	-4.100^{a}	-4.207^{a}	-4.138^{a}	-4.003^{a}	-4.613^{a}	-3.659^{a}	-4.098^{a}
EUROSTOXX50 (returns)	31.068^{a}	4.055	3.402	8.336	11.189	8.601	15.328
cond. volatility of EUROSTOXX50 returns	534.693	1089.408^{a}	685.118^{c}	943.504^{b}	998.320^{a}	1026.443^{a}	789.804^{b}
EURIBOR (level)	-0.138	-0.140	-0.127	-0.148	-0.026	-0.159	-0.154
USA 10y T.N. yield changes	120.311^{a}	84.020^{a}	55.205^{a}	73.672^{a}	74.32333^{a}	75.43978^{a}	68.440^{a}
EUR-USD returns	-20.364	-132.768^{a}	-116.402^{a}	-112.486^{a}	-117.266^{a}	-126.023^{a}	-102.325^{a}
Crisis period 1	0.536	-0.543	0.409	-0.319	0.255	-0.653^{a}	0.314
Crisis period 2	-0.770	0.827^{c}	2.172^{a}	1.311	1.345^{a}	0.821^{b}	1.433^{a}
Log likelihood	-635.01	-754.90	-708.00	-722.17	-764.40	-677.96	-736.98
LR chi (14)	544.61	350.08	333.64	353.66	340.20	293.40	356.52
Prob.>chi2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo-R^2 (McFadden)	0.300	0.188	0.191	0.197	0.182	0.178	0.195

Notes: See notes for Table 4a.

I find that the regression coefficients of the EUROSTOXX50 returns, U.S. 10-year Treasury note yield changes, and for some pairs of sovereign bond markets, the EUR-USD returns and time dummies are significantly different from zero. From the data in Table 4a, it follows that for the sovereign bond markets of France and Germany, a one unit (i.e., a 1%) increase in the EUROSTOXX50 returns is associated with a 0.536[13] increase in the relative log odds of outcome 2 (i.e., exceedance in one of the sovereign bond markets) versus outcome 1 (i.e., no exceedance in any of the two observed markets). One can also see a 0.578 increase in the relative log odds of outcome 3 (i.e., co-exceedance) versus outcome 1. Similarly, a one unit increase (i.e., 1%) in U.S. 10-year Treasury note yields is associated with a 1.489 (1.311) increase in the relative log odds of outcome 2 (outcome 3). The pseudo-R^2 is between 0.18 and 0.30. For economic interpretation of the regression coefficients, however, one must calculate the marginal effects of the estimated coefficients ([12]). The marginal effects and probabilities of outcomes are reported in Tables 5a and 5b.

Turning first to estimated probabilities of outcomes, I find that the probabilities of joint extreme yield increases in the pair-wise investigated sovereign bond markets range between 0.0205 (or around 2%), for the sovereign bond markets of Germany–Portugal, and 0.0397 (or around 4%), for the sovereign bond markets of France-Germany, when not controlling for the covariates (see reported Probabilities 1 in Tables 5a and 5b).[14] As noted, to separate contagion from interdependence, it is important to control for common and regional factors that impact all countries simultaneously. Evidently, this reduces the probabilities of observing outcome 3 (i.e., contagion) because it now ranges between 0.0083 for the sovereign bond markets of Germany–Portugal and 0.0168 for the sovereign bond markets of Ireland–Portugal (see Probabilities 2, reported in Tables 5a and

[13] 0.01*53.650, as in the data a 1% is expressed as 0.01.

[14] If the outcomes were independent, then the probabilities of co-exceedances between all sovereign bond markets investigated pair-wise would be $0.05^2 = 0.0025$.

5b). The probability of observing no extreme yield moves in any of the two observed sovereign bond markets is the highest for France–Germany (0.9798) and lowest for Italy–Portugal (0.958). The probability of extreme yield movement in just one of the observed markets in the pair is the lowest for the bond markets of France–Germany (0.0088) and the highest for the markets of Ireland–Italy (0.0293). The results indicate that the sovereign bond yield dynamics of France and Germany were not only the most correlated (see Table 2) of all the markets investigated, but also had a very similar time path of extreme yield dynamics during the entire observed period.

Looking at specific covariates, EUROSTOXX50 returns, U.S. Treasury note yields, and EUR-USD exchange rate significantly impact the probability of extreme yield increases in the Eurozone sovereign bond markets. While positive EUROSTOXX50 returns and increased yields of U.S. Treasury notes increase the probability of extreme sovereign bond yield across the Eurozone, the increase in the EUR-USD exchange rate (i.e., appreciation of the EUR against the USD) significantly reduces the probability. The conditional volatility in the Eurozone stock markets and the money market interest rate do not significantly impact the probability of extreme yield movements in the Eurozone's sovereign bond markets.

The responsiveness of the dependent (i.e., co-exceedance) variable to shocks in the Eurozone stock markets is not uniform across the sovereign bond markets. For example, a 1% increase in EUROSTOXX50 returns significantly (at the 5% level) increases the probability of extreme increase in bond yields in either the sovereign bond markets of France or Germany of 0.0046 (or 0.46%) and a probability of observing a contagion (i.e., a simultaneous extreme yield increase in both markets) of 0.0051 (or 0.51%). A similar response to shocks in the Eurozone stock markets are observed for the bond markets of Germany–Portugal and Germany–Spain. In other pair-wise investigated sovereign bond markets, only the probability of outcome 2 or outcome 3 significantly increased; otherwise, the impact is insignificant (the latter can be noticed for the sovereign bond markets of Ireland–Italy, Ireland–Portugal, and Ireland–Spain). The increased volatility in the Eurozone stock markets significantly increases the probability of simultaneous extreme yield dynamics in the sovereign bond markets of France–Italy, Ireland–Spain, Italy–Portugal, Italy–Spain, and Portugal–Spain.

Table 5a

Marginal effects and probabilities of outcomes for particular pair-wise observed sovereign bond markets

	Fra-Ger	Fra-Ire	Fra-Ita	Fra-Por	Fra-Spa	Ger-Ire	Ger-Ita	Ger-Por
Outcome 2								
EUROSTOXX50 (returns)	0.4648[a]	0.3220	0.1115	0.4066[c]	0.2337[c]	0.4723[a]	0.3106[c]	0.5340[a]
cond. volatility of EUROSTOXX50 returns	1.2866	8.6072	4.5548	6.5728	7.3644	3.1658	3.4873	4.9371

EURIBOR (level)	-0.0002	-0.0057^{b}	0.00264	-0.0056^{c}	-0.0027	-0.0012	0.0023	-0.0040
USA 10y T.N. yield changes	0.7805^{a}	1.3656^{a}	0.8401^{a}	1.5341^{a}	0.6034^{a}	0.9815^{a}	1.4475^{a}	1.2937^{a}
EUR-USD returns	-0.2039	-2.0148^{b}	-0.4195	-2.4714^{a}	-1.4202^{a}	-0.7678^{b}	-0.3808	-1.2619^{a}
Crisis period 1	0.0048	0.0210	0.0681^{a}	0.0344^{b}	0.0163	0.0780^{a}	0.0609^{a}	0.0650^{a}
Crisis period 2	0.0309^{c}	0.1535^{a}	0.2218^{a}	0.1588^{a}	0.1306^{a}	0.2823^{a}	0.2184^{a}	0.2194^{a}
Outcome 3								
EUROSTOXX50 (returns)	0.5113^{a}	0.2028^{b}	0.2267^{b}	0.2287^{b}	0.3247^{a}	0.1744^{c}	0.2325^{b}	0.2312^{b}
cond. volatility of EUROSTOXX50 returns	1.5400	8.2095^{c}	10.1188^{b}	6.5010	7.3173	6.8400^{c}	8.2398^{b}	4.9732
EURIBOR (level)	-0.0007	-0.0003	-0.0013	-0.0006	-0.0013	-0.0008	-0.0010	-0.0004
USA 10y T.N. yield changes	1.4710^{a}	1.0412^{a}	1.357^{a}	0.9150^{a}	1.4018^{a}	0.9961^{a}	1.1015^{a}	0.8806^{a}
EUR-USD returns	0.0473	-0.4890^{c}	-0.7773^{a}	-0.3831	-0.4321	-0.4049^{c}	-0.4821^{b}	-0.3470
Crisis period 1	-0.0008	-0.0037	-0.0094^{a}	-0.0009	-0.0061^{c}	-0.0049^{b}	-0.0056^{b}	-0.0003
Crisis period 2	0.0103	0.0019	-0.0088^{a}	0.0001	-0.0080^{b}	-0.0024	-0.0065^{b}	-0.0001
Probabilities 1								
Outcome 1	0.9395	0.9237	0.9324	0.9213	0.9321	0.9227	0.9267	0.9203
Outcome 2	0.0208	0.0525	0.0350	0.0572	0.0356	0.0545	0.0464	0.0592
Outcome 3	0.0397	0.0239	0.0326	0.0215	0.0323	0.0229	0.0269	0.0205
Probabilities 2								
Outcome 1	0.9798	0.9650	0.9739	0.9649	0.9753	0.9720	0.9708	0.9701
Outcome 2	0.0088	0.0252	0.0147	0.0264	0.0122	0.0192	0.0200	0.0216
Outcome 3	0.0114	0.0099	0.0115	0.0088	0.0126	0.0088	0.0092	0.0083

Notes: Probabilities 1 are probabilities of outcomes when one does not control for covariates. Probabilities 2 are probabilities of outcomes after controlling for the covariates and are calculated by Equation (4). ^{a}/^{b}/^{c} denote the 1%, 5%, 10% significance of the rejection of the null hypothesis that the marginal effect of the covariate is equal to 0 based on z-statistics. The reported marginal effects of the time dummy covariates (Crisis period 1, Crisis period 2) show by how much the probability of observing outcome 2 (outcome 3) increases when the value of the time dummy variable changes from 0 to 1.

A significant impact on the probability of extreme yield increases in all the Eurozone's sovereign markets is exerted by the yield dynamics of the U.S. Treasury notes. For instance, a 1% increase in the yields of U.S. Treasury notes increases the probability of an extreme increase in the yields in the sovereign bonds of either France or Germany by 0.78% and the probability of contagion between the bond markets of France and Germany by 1.47%.

Appreciation of the euro against the U.S. dollar reduces the probability of extreme increases in the yields of Eurozone sovereign bond markets, except in the markets of France–Germany. The negative relationship between the exchange rate and sovereign bond yields dynamics can be explained as follows. Increased EUR-USD exchange rate increases the demand for Eurozone bonds as the expected return on the euro denominated bond investment increases. Increased demand for bonds, in turn, increases their prices and reduces their yields. The marginal effect of covariate is significantly different from zero across all the pair-wise observed sovereign bond markets.

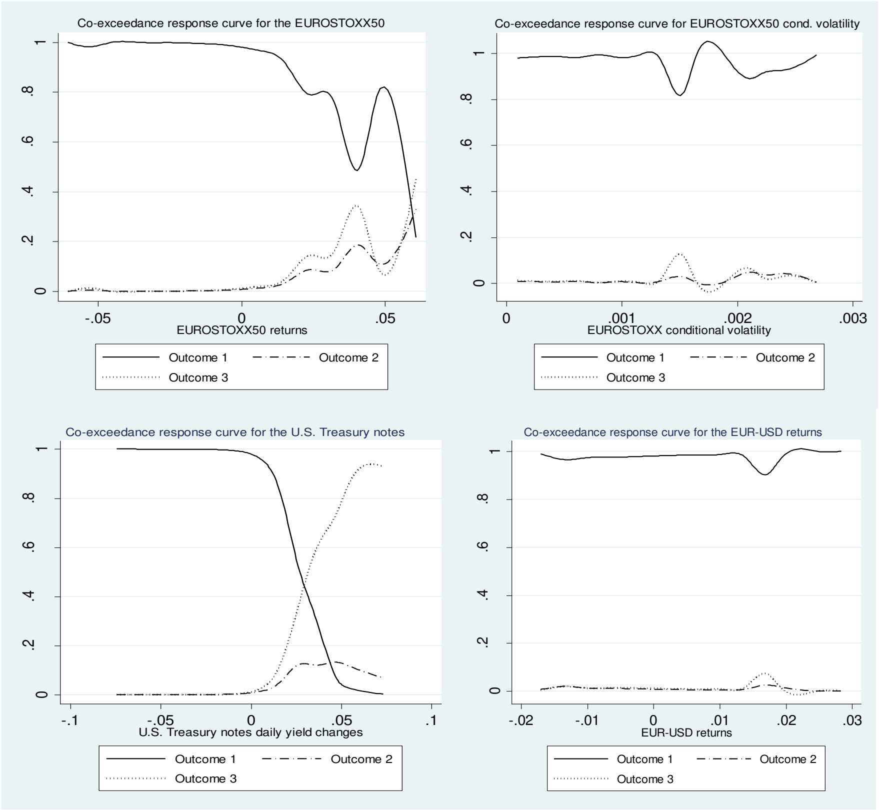
Table 5b
Marginal effects and probabilities of outcomes for particular pair-wise observed sovereign bond markets

	Ger-Spa	Ire-Ita	Ire-Por	Ire-Spa	Ita-Por	Ita-Spa	Por-Spa
Outcome 2							
EUROSTOXX50 (returns)	0.3621^{a}	0.0589	0.3113^{c}	0.1116	0.0657	-0.01339	0.1066
cond. volatility of EUROSTOXX50 returns	3.9344	12.5523	4.6130	1.3537	7.4200	1.8063	5.4658
EURIBOR (level)	-0.0008	0.0002	-0.0029	-0.0020	-0.0071^{b}	0.0023	-0.0043
USA 10y T.N. yield changes	0.7073^{a}	0.9514^{a}	0.4223^{a}	0.4957^{a}	1.1792^{a}	0.5924^{a}	.7338^{a}
EUR-USD returns	-1.0622^{a}	-1.4973^{a}	-1.2112^{a}	-1.5689^{a}	-2.4025^{a}	-0.2888	-2.4376^{a}
Crisis period 1	0.0493^{b}	0.0464^{b}	0.0581^{a}	0.0522^{b}	0.0229	0.0650^{b}	0.0373^{b}
Crisis period 2	0.2243^{a}	0.221^{a}	0.1181^{a}	0.1809^{a}	0.1172^{a}	0.1412^{a}	0.1425^{a}
Outcome 3							
EUROSTOXX50 (returns)	0.3053^{a}	0.0482	0.0508	0.1262	0.1427	0.1658	0.2234
cond. volatility of EUROSTOXX50 returns	5.2769	12.9953	11.2122^{c}	14.4642^{b}	12.7120^{b}	19.714^{a}	11.511^{b}
EURIBOR (level)	-0.0014	-0.00169	-0.0021	-0.0022	-0.0002	-0.0031	-0.00219
USA 10y T.N. yield changes	1.1891^{a}	1.0025^{a}	0.9026^{a}	1.1232^{a}	.9380^{a}	1.4398^{a}	0.9935^{a}
EUR-USD returns	-0.1917	-1.5842^{a}	-1.8977^{a}	-1.7020^{a}	-1.4727^{a}	-2.4191^{a}	-1.4650^{a}
Crisis period 1	-0.0048^{c}	-0.0059^{c}	0.0064	-0.0050	0.0032	-0.0109^{b}	0.0044
Crisis period 2	-0.0071^{a}	0.0083	0.0786^{a}	0.0259^{c}	0.0254^{c}	0.0162	0.0310^{b}
Probabilities 1							
Outcome 1	0.9287	0.9257	0.9324	0.9297	0.9250	0.9371	0.9277
Outcome 2	0.0424	0.0484	0.0350	0.0403	0.0498	0.0256	0.0444
Outcome 3	0.0289	0.0259	0.0326	0.0299	0.0252	0.0373	0.0279
Probabilities 2							
Outcome 1	0.9763	0.9585	0.9659	0.9640	0.9580	0.9663	0.9626
Outcome 2	0.0137	0.0293	0.0174	0.0204	0.0289	0.0141	0.0225
Outcome 3	0.0100	0.0122	0.0168	0.0156	0.0130	0.0196	0.0149

Notes: See notes for Table 5a.

To answer the question of whether the probability of contagion increased during the global financial crisis and the Eurozone debt crisis, the marginal effects of the dummy variables of Crisis period 1 and Crisis period 2 must be analyzed. The time-dummy covariates significantly impact the probability of (co-)exceedances across the pair-wise observed markets, except the sovereign bond markets of France–Germany. The probability of observing exceedance (i.e., outcome 2) during the Eurozone debt crisis increased dramatically compared to the pre-crisis period. For example, the probability of extreme upside movement in sovereign bond yields increased by 0.28 (or 28%) compared to the pre-crisis period when simultaneously analyzing Germany's and Ireland's exceedance time series. The probability of observing co-exceedance (i.e., outcome 3) of extreme positive bond yield changes during the Eurozone debt crisis increased for the sovereign bond markets of Ireland–Portugal and for Portugal–Spain, whereas the probability reduced for the markets in France–Spain and Germany–Spain. The estimates of the marginal effects of the time dummies indicate less synchronous extreme yield dynamics across the Eurozone sovereign bond markets during the global financial crisis and especially during the Eurozone debt crisis compared to the pre-crisis periods.

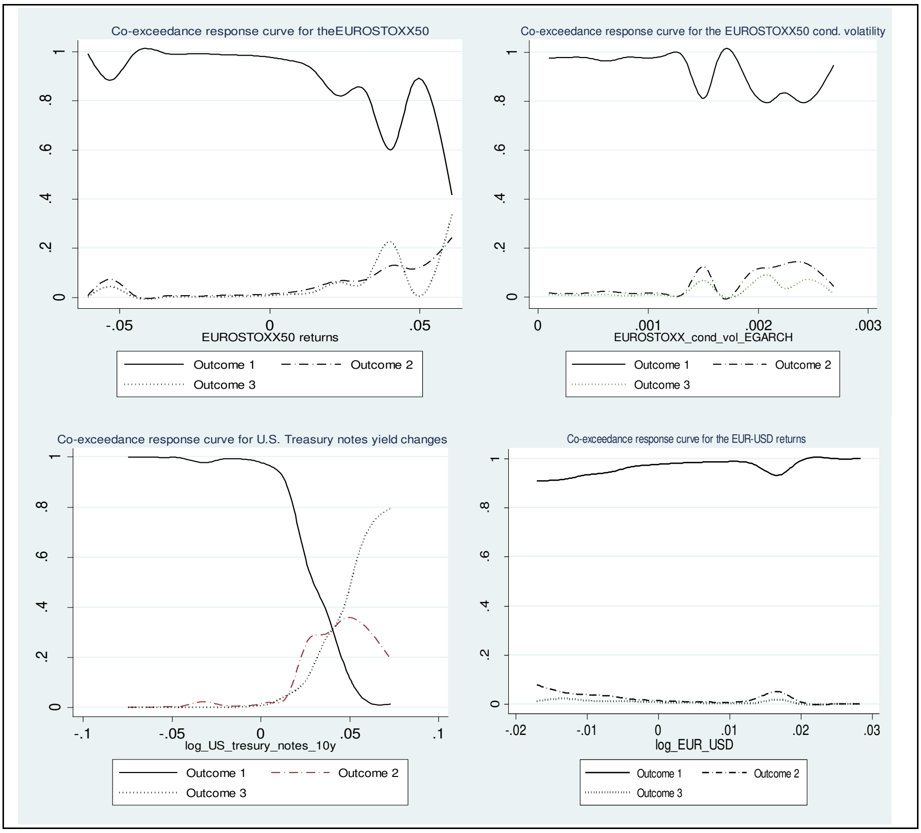
Responses of the probability estimates to the full range of values associated with different covariates are computed and graphically presented in Figures 2a (for the France–Germany) and 2b (for Germany–Portugal).



Notes: Marginal effects are calculated by Equation (5).

Figure 2a
Co-exceedance response curves of the sovereign bond markets of France-Germany to changes in covariates

The probability of (co-)exceedance in bond markets clearly increases with the stock market returns, but it does so nonlinearly. Stock market volatility increases the probability of (co-)exceedance only after it reaches some threshold. From Figures 2a and 2b, this is when conditional variance exceeds 0.0015 a day. The U.S. Treasury note yield increases highly increase the probability of co-exceedances in the Eurozone bond markets when yields in the U.S. bond market increase by more than 0.02 (2%) a day, whereas the EUR-USD returns increase the probability of co-exceedance between the bond markets of Germany-Portugal when the exchange rate falls by more than 1% a day.



Notes: Marginal effects are calculated by Equation (5).

Figure 2b

Co-exceedance response curves of the German-Portugal's sovereign bond markets to changes in covariates

Conclusion

In the present paper, a multinomial logistic model was applied to analyze contagion between six Eurozone countries based on a measure of joint occurrences of large yield changes (i.e., co-exceedances), controlling for developments in common and regional factors that affect all sovereign bond market simultaneously. I found that Eurozone stock markets (EUROSTOXX50 returns), U.S. Treasury note yields and the EUR-USD exchange rate significantly impact the probability of extreme yield moves in the Eurozone sovereign bond markets. Whereas positive EUROSTOXX50 returns and positive U.S. Treasury note yield moves increase the probability of extreme positive sovereign bond yields in the Eurozone sovereign bond markets, appreciation of the euro against the U.S. dollar significantly reduces the probability. The conditional volatility in the Eurozone stock markets and the money market interest rate do not significantly impact the probability of extreme yield movements in the Eurozone's sovereign bond markets.

The probability of observing exceedance during the Eurozone debt crisis increased dramatically compared to the pre-crisis periods. The probability of observing co-exceedance of extreme positive bond yield changes during the Eurozone debt crisis increased for the sovereign bond markets of Ireland–Portugal and Portugal-Spain, whereas the probability reduced for the markets in France–Spain and Germany-Spain. The results indicate a less synchronous extreme yield dynamic across the Eurozone sovereign bond markets during the global financial crisis and especially during the Eurozone debt crisis compared to a pre-crisis period.

The results of the present study might be of interest for policymakers, central banks, and investors in financial markets. Through contagion analysis, one gains knowledge of whether a shock in one segment of a national financial market is transmitted across markets via channels that appear only during turbulent periods or whether these shocks are transmitted via channels or inter-linkages that exist in all states of the world (non-crisis as well as crisis periods).

References

[1] Bae, K. H, Karolyi, G. A. and Stulz, R. M. (2003) 'A New Approach to Measuring Financial Contagion', *Review of Financial Studies,* 16(3), pp. 717-763

[2] Baur, D. G. and Lucey, B. M., (2009) 'Flights and Contagion - An Empirical Analysis of Stock-Bond Correlations, *Journal of Financial Stability* 5(4), pp. 339-352

[3] Baur, D. and Schulze, N. (2005) 'Coexceedances in Financial Markets - A Quantile Regression Analysis of Contagion', *Emerging Markets Review*, pp. 5(1), 21-43

[4] Corsetti, G., Pericoli, M. and Sbracia, M. (2001) 'Correlation Analysis of Financial Contagion: What One Should Know Before Running a Test'. *Economic Growth Center, Yale University Discussion Paper* No. 822

[5] Dornbusch, R., Park, Y. C. and Claessens, S. (2001) '*Contagion: Why crises spread and how this can be stopped*', in Claessens, S., and Forbes, K. (Eds.), International Financial Contagion, Kluwer Academic Publishers, Boston, pp. 19-42

[6] Dungey, M., Fry, R., Gonazlés-Hermosillo, B. and Martin, V. L. (2005) 'Empirical Modelling of Contagion: A review of Methodologies', *Quantitative Finance* 5(1), pp. 9-24

[7] Dungey, M., and Martin, V. L. (2007) 'Unraveling Financial Market Linkages during Crises', *Journal of Applied Econometrics*, 22(1), pp. 89-119

[8] Durré, A. and Giot, P. (2005) ‘An International Analysis of Earnings, Stock Prices and Bond Yields’. *European central bank working paper* No. 515

[9] Forbes, K. J. and Rigobon, F. (2001) ‘*Measuring contagion: Conceptual and empirical issues*’, in Claessens, S., and Forbes, K. (Eds.), International Financial Contagion, Kluwer Academic Publishers, Boston, 43-66

[10] Forbes, K. and Rigobon, R. (2002) ‘No Contagion, Only Interdependence: Measuring Stock Market Co-movements’, *Journal of Finance* 57(5), pp. 2223-61

[11] Gravelle, T., Kichian M. and Morley, J. (2006) ‘Detecting Shift Contagion in Currency and Bond Markets’, *Computing in Economics and Finance 2002*, 58. Society for Computational Economics

[12] Greene, W. H. (2003) ‘*Econometric Analysis*’, 5th ed., Prentice Hall, New Jersey

[13] Kim, S. and In, F. (2007) ‘On the Relationship in Stock Prices and Bond Yields in the G7 Countries: Wavelet analysis’, *Journal of International Financial Markets, Institutions and Money*, 17(2), pp. 167-179

[14] Longin, F. M. and Solnik, B., (1995) ‘Is the Correlation in International Equity Returns Constant: 1970-1990?’, *Journal of International Money and Finance*, 14(1), pp. 3-26

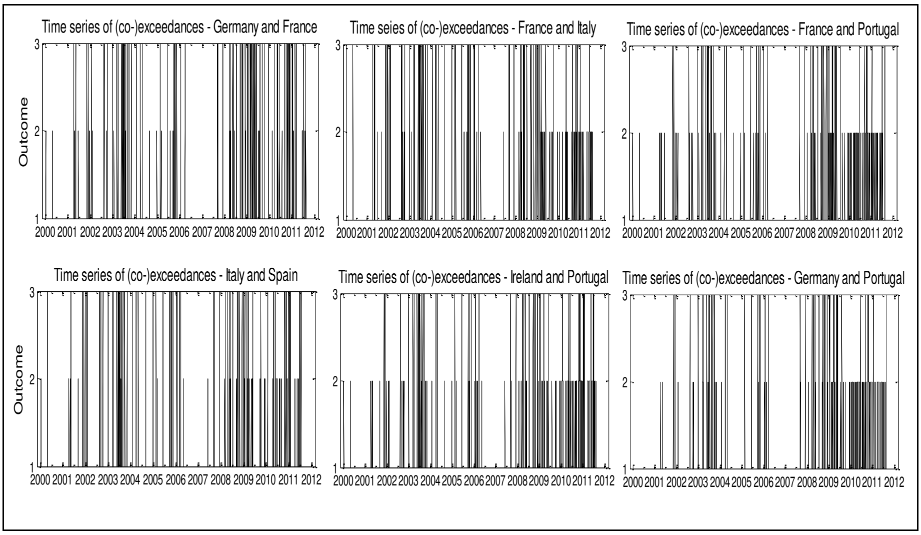
[15] McFadden, P. (1974) ‘The Measurement of Urban Travel Demand’, *Journal of Public Economics*, 3(4), pp. 3303-3328

[16] Metiu, N. (2011) ‘Financial Contagion in Developed Sovereign Bond Markets’. *METEOR – Maastricht Research School of Economics of Technology and Organization, Research Memoranda* 004

[17] Nelson, D. B. (1991) ‘Conditional Heteroskedasticity in Asset Returns: A New Approach’, *Econometrica*, 59(2), pp. 347-370

[18] Pericoli, M. and Sbracia, M. (2003) ‘A Primer on Financial Contagion’, *Journal of Economic Surveys*, 17(4), pp. 571-608

Appendix



Notes: Only the time series of (co-)exceedances of yield changes of pair-wise observed sovereign bond markets with the highest (first two columns of plots) and the lowest (the last column of plots) co-exceedance counts are presented. Outcome 1 presents the occurrence of category 1 (i.e., no exceedance in any of the sovereign bond market yield changes); outcome 2 presents the occurrence of category 2 (i.e., exceedance in one of the sovereign bond market yield changes); and outcome 3 presents the occurrence of category 3 (i.e., co-exceedance of upper 5% yield changes in both of the national sovereign bond markets).

Figure 1

Time series of (co-)exceedances in yield changes for several of the pair-wise observed sovereign bond markets

Loading Surface in the Course of Mechanical-Thermal Treatment and Steady-State Creep of Metals

Andrew Rusinko

Óbuda University
Népszínház utca 8, H-1081 Budapest, Hungary
E-mail: rusinko.endre@bgk.uni-obuda.hu

Abstract: Kinetics of the loading surface of a material gives precious information on the level of the hardening of the material. This paper is concerned with the evolution of the loading surface during successive actions, such as: (i) plastic deformation, (ii) annealing of the pre-strained specimen, and (iii) secondary creep of the treated material. The analysis of the loading surface is carried out in terms of the synthetic theory of irrecoverable deformation.

Keywords: loading surface; mechanical-thermal treatment; creep and plastic strain; synthetic theory of irrecoverable deformation

1 Introduction

Numerous experiments testify that mechanical-thermal treatment (MTT) is an effective tool to improve the strength of metals [2, 4, 9, 10]. MTT involves (i) plastic deformation of a specimen at room temperature (T_0), e.g. in uniaxial tension (we give the acting stress symbol σ_{x0} and we will mark as ε_{x0} the plastic strain induced by σ_{x0}), and (ii) annealing of the cold-worked specimen in unloaded state (let us denote the annealing temperature and duration as T_1 and t_1, respectively). As seen in Figures 1 and 2, if we subject the treated specimens to creep with stress σ_x and temperature T_2, the creep rate ($\dot{\varepsilon}_x$) is not a monotonous function of the plastic pre-strain ε_{x0} (with the proviso that the values of σ_x, T_2, T_1, and t_1 are unchangeable). Figure 1 demonstrates the steady-state creep rate $\dot{\varepsilon}_x$ ($\sigma_x = 25\,\text{MPa}, T_2 = 700°\text{C}$) of Ni+1.18% alloy against plastic pre-strain at room temperature $\dot{\varepsilon}_{x0}$ developed in the course of preliminary MTT (the annealing temperature and duration $T_1 = T_2 = 700°\text{C}$ and $t_1 = 1\,\text{hour}$, respectively). Figure 2 shows the steady-state creep rate ($\sigma_x = 15\,\text{MPa}$, $T_2 = 500°\text{C}$) of copper as a function of plastic pre-strain at room temperature developed in MTT ($T_1 = T_2 = 500°\text{C}$, $t_1 = 1\,\text{hour}$). In these figures, the points are the experiment [2,4], and the solid line is the analytical curve.

According to Figures 1 and 2, there exists an optimal level of plastic pre-straining after which (with intermediate annealing) the rate of stationary creep is minimal. The existence of different types of behavior of the tested specimens subjected to creep is connected with distinctions in the initial structure of the material formed as a result of plastic pre-straining and annealing. Indeed, the structure affects the intensity of the processes of polygonization and recrystallization, which control the rate of stationary creep.

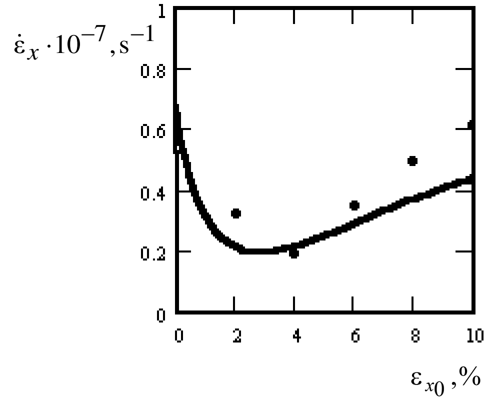


Figure 1

Steady-state creep rate of Ni+1.18% alloy against plastic pre-strain

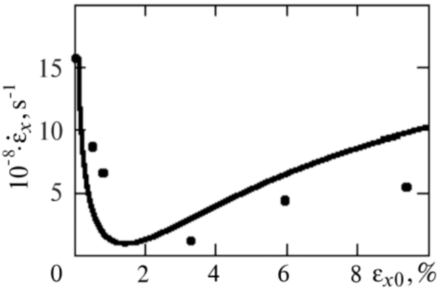


Figure 2

Dependence of the steady-state creep rate of copper on the level of preliminarily induced plastic strains

Another positive effect of MTT is an increase in the duration of the secondary creep [2,4], which can be concluded from cavities nucleation and development [3].

Due to a plastic strain, the initial (relatively perfect) crystals suffer fragmentation and the sizes and orientation of fragments depend on the level of strains. The boundaries between the fragments form a three-dimensional grid of dislocations which can be regarded as a pileup of dislocations. Subsequent annealing ambiguously affects the preliminarily formed dislocation structure promoting the initiation of the thermally controlled processes of polygonization and recrystallization. The course of one of these processes depends on the level of pre-straining and the temperature of annealing [5-7, 12].

In the case of annealing of an insignificantly cold-hardened material, we observe the redistribution of dislocations of the same sign in the form of rearrangement into vertical walls. As a result, poorly formed cells appearing as a result of plastic deformation become completely surrounded with low-angle boundaries and gradually turn into well-formed subgrains in the body of which the density of dislocations is lower than in the deformed matrix. The higher (to a certain degree) the value of plastic pre-straining, the greater the number of appearing subgrains, i.e., the higher the intensity of polygonization. The polygonized structure formed in the process of preliminary TMT decreases the level of sliding (both coarse and fine) in testing for creep. This fact can be explained by the restriction of the free path of dislocations imposed by the preliminarily formed grid of subgrain boundaries.

In the case of annealing of a material with relatively high plastic strains, the density of dislocations built in the subgrain boundaries increases and, hence, the angle of their mutual orientation also increases. It is known that subgrains with large-angle boundaries play the role of centers of recrystallization, i.e., of the formation and growth of grains with more perfect structure. In the course of recrystallization, the resistance of the metal to plastic deformation significantly decreases since the rapid migration of the boundaries intensely “cleans” the deformed matrix, which facilitates the motion of dislocations under conditions of creep and increases the rate of stationary creep as compared with its optimal value. Thus, the optimal degree of plastic pre-straining should be chosen to avoid the possibility of intense recrystallization.

Classic theories such as ageing (time-hardening) theory, flow theory, and strain-hardening theory [1] are incapable of the modeling of the dependence $\dot{\varepsilon}_x = f\left(\varepsilon_{x0}\right)$ due to the fact that, in terms of these theories, the creep rate is related to the acting stresses and the loading-prehistory is ignored. This fact motivates the author to model the effect of MTT utilizing a more efficient model able to embrace a wider circle of processes and their interplay: the synthetic theory of irrecoverable deformation (ST).

2 Mathematical Apparatus: Synthetic Theory

The analytical description of the dependence of steady-state creep rate $\dot{\varepsilon}_x$ on the plastic pre-strain ε_{x_0} in the course of MTT is carried out in terms of the synthetic theory of irrecoverable deformation [8-11]. The yield limit of a material in the three-dimensional subspace ($\mathbf{R}^3$) of five-dimensional stress deviator space has the shape of a sphere, which correspond to the von-Mises yield criterion:

$$S_1^2 + S_2^2 + S_3^2 = 2/3\sigma_T^2 \tag{1}$$

where σ_T, depending on the problem considered, is the yield limit (σ_S) or the creep limit (σ_S) of the material; S_i ($i = 1,2,3$) are coordinate axes in $\mathbf{R}^3$.

In terms of ST, the yield/loading surface is constructed as the inner envelope of tangent planes. In the virgin state, sphere (1) is the inner envelope of tangent planes, which are equidistant to the origin of coordinate in all directions. A loading is presented by a stress-vector, $\vec{\mathbf{S}}$, whose components are defined in [8, 11]. As the stress-vector grows, it translates on its endpoint tangent planes (the orientation of tangent plane does not vary during the motion). The planes which are not reached by $\vec{\mathbf{S}}$ remain stationary. The displacement of a plane symbolizes the increment in plastic deformation (plastic slip) within an appropriate slip system at a point of a body. The plastic strain developed within one slip system defines a microlevel of deformation. The total strain (macrostrain) is calculated by the summation (threefold integrals) of the microstrains occurring in activated slip systems.

Consider the case of uniaxial tension when the components of $\vec{\mathbf{S}}$ are $S_1 = \sqrt{2/3}\sigma_x$, $S_2 = 0$, $S_3 = 0$ [8]. As $|\vec{\mathbf{S}}| = \sqrt{2/3}\sigma_S$, there is only one plane tangential to the sphere (1) reached by the vector $\vec{\mathbf{S}}$, and it is perpendicular to the $\vec{\mathbf{S}}$. During further elongation of the $\vec{\mathbf{S}}$, new planes become located on the endpoint of the stress-vector and – following the rule that a loading surface is the inner envelope of planes – the loading surface take the shape of a cone symmetric relative to the S_1-axis, and its generator is constituted of the boundary tangent planes reached by the stress-vector. This cone goes over to the initial sphere in the directions where tangent planes remain immovable.

The plastic strain component in uniaxial tension, e_1, is calculated as [8-11]

$$e_1 = \frac{1}{r}\int_{\alpha}\int_{\beta}\int_{\lambda}\varphi_N \sin\beta\cos\beta\cos\lambda d\beta d\lambda, \quad e_2 = 0,\ e_3 = 0, \tag{2}$$

where φ_N is referred to as irrecoverable strain intensity (index N stands for the vector normal to the tangent plane, which gives the orientation of the plane), which is defined by the following differential equation

$$d\psi_N = rd\varphi_N - K\psi_N dt , \tag{3}$$

where φ_N is called defects intensity; dt is the time differential, r is the constant of material, and K is a function of the homological temperature and current stresses:

$$K = K_1 \exp(K_2\Theta)\left(\sqrt{2/3}\sigma_x\right)^{K_3} , \tag{4}$$

where K_i ($i = 1,2,3$) are the material constants.

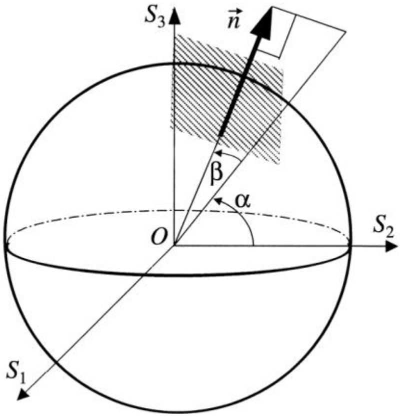


Figure 3
Orientation of the normal $\vec{\mathbf{n}}$ in the subspace $\mathbf{R}^3$

The defects intensity, according to [8, 9], is

$$\psi_N = H_N^2 - 2/3\sigma_T{}^2 \ , \quad H_N = \vec{\mathbf{S}} \cdot \vec{\mathbf{N}} = \vec{\mathbf{S}} \cdot \vec{\mathbf{n}} \cos\lambda = 2/3\sigma_x \sin\beta\cos\lambda , \tag{5}$$

where H_N ($i = 1,2,3$) is a distance to the plane reached by stress-vector; angle β gives the orientation of a plane in $\mathbf{R}^3$ (the role of angle λ, which is immaterial within this article, can be found in [8]). In general, the orientation of the normal in $\mathbf{R}^3$, $\vec{\mathbf{n}}(\sin\beta, \cos\alpha\cos\beta, \sin\alpha\cos\beta)$, is shown in Figure 3. The magnitude of H_N shows the degree of the hardening of material. Actually, the greater the H_N, the greater the stress-vector needed to reach the plane.

When calculating “immediate” plastic deformation ($dt = 0$), the formula takes the form

$$\psi_N = r\varphi_N . \tag{6}$$

For the case of secondary creep ($\dot{\psi}_N = 0$), from (3) we have

$$\dot{\varphi}_N = K\psi_N / r. \tag{7}$$

The steady-state creep rate strain component is expressed by the relationship where the integrand is $\dot{\varphi}_N$.

If annealing a work-hardened specimen in an unloaded state ($d\varphi = 0$), the formula yields

$$d\psi_N = -K\psi_N dt. \tag{8}$$

The solution of differential equation (8) is

$$\psi_N = \psi_{N_0} \exp(-Kt),$$

where ψ_{N_0} is the defects intensity in the work-hardened material. Therefore, on account of formula (5), the distance to the planes after the plastic deformation and annealing is

$$H_N^2 = \psi_{N_0} \exp(-Kt) + 2/3\sigma_T^2. \tag{9}$$

Formula (9) expresses the motion of the planes toward the origin of coordinates. This motion for each plane will terminate as it touches the sphere (1).

3 The Generalization of the Synthetic Theory to the Case of MTT

To evaluate the steady-state creep rate of a metal after MTT, we replace formulae (5) and (4) by

$$\psi_{M_N} = H_N^2 - H_{T_N}^2, \tag{10}$$

$$K_M = f(\Theta, H_{\max}) = K_1 \exp(K_2\Theta) H_{\max}^{K_3}, \tag{11}$$

where H_{T_N} is the distance to a plane after MTT, which characterizes the thermal stability of the polygonization structure against the recrystallization during creep. In the absence of MTT, $H_{T_N} = \sqrt{2/3}\sigma_P$ and the relation for ψ_{M_N} from (10) degenerates to (5). In formula (10), $H_{\max}$ is the maximal distance to planes for the whole loading history at a given temperature.

3.1 Tension of Specimen at Room Temperature

The plastic strain component, e_{1_0}, is calculated as [8]

$$e_{1_0} = a_0 \Phi\left(\sin\beta_{1_0}\right), \; a_0 = \pi\sigma_S^2 / (9r), \; \sin\beta_{1_0} = \sigma_S / \sigma_{x_0}, \tag{12}$$

$$\Phi(\xi) = \left(2\sqrt{1-\xi^2} - 5\xi^2\sqrt{1-\xi^2} + 3\xi^4 \ln\frac{1+\sqrt{1-\xi^2}}{\xi} \right)\frac{1}{\xi^2}. \tag{13}$$

The loading surface is shown in Figure 4a. The value of $H_{\max}$ in the plastic loading can be obtained from (5) at $\beta = \pi/2$, $\lambda = 0$.

3.2 Annealing of Deformed Specimen

Consider the case when the annealing temperature coincides with the temperature of the following creep: $T_1 = T_2$.

As seen from (5), the heating of the specimen to the temperature T_1 results in a decrease of H_N due to a drop in the value of σ_T caused by the temperature gradient. Since the temperatures of annealing and creep are assumed to be equal, the effect of MTT on the steady-state creep rate can be revealed by studying the positions of tangent planes relative to the value of σ_P at $T = T_1$. If we ignore an immaterial duration of the heating from T_0 to T_1, the decrease in H_N can be assumed to be of a step-wise nature, which symbolizes the step-wise motions of planes toward the origin of coordinates. This fact means that the values of angles λ_{1_0} and β_{1_0} remain unchangeable during annealing. Therefore, formula (9) at $t = 0$ takes the form

$$H_N^2 = \frac{2}{3}\begin{cases} \left[\left(\sigma_{x_0}\sin\beta\cos\lambda\right)^2 - \sigma_S^2\right] + \sigma_P^2, & \beta_{1_0} \le \beta \le \pi/2, \quad 0 < \lambda < \lambda_{1_0} \\ \sigma_P^2 & -\pi/2 < \beta < \beta_{1_0}, \quad \lambda_{1_0} < \lambda \end{cases}. \tag{14}$$

Consider the distance from the origin of coordinates to the point of intersection of tangent plane with S_1-axis, $L(\beta, \lambda, t)$. For simplicity, while not distorting the result, we study the value of L at $\lambda = 0$. From Figure 4a it follows that

$$L(\beta, t) = H_N / \sin\beta. \tag{15}$$

In active loading, $L(\beta) = const = \sqrt{2/3}\sigma_{x_0}$ for $\beta_{1_0} \le \beta \le \pi/2$. The value of L due to the temperature decrease, $\Delta T = T_1 - T_0$, according to (14) and (15), is

$$[L(\beta, t=0)]^2 = \frac{2}{3}\left[\sigma_{x_0}^2 - \left(\sigma_S^2 - \sigma_P^2\right)\Big/\sin^2\beta\right], \qquad \beta_{1_0} \le \beta \le \pi/2. \tag{16}$$

As seen from formula (16), the loading surface maintains its shape but the tangent planes from the range $\beta_{1_0} < \beta \le \pi/2$ are not on the cone tip (Figure 4b).

During annealing, according to (9), (11), and (15), we have

$$H_N^2 = \frac{2}{3}\left[\left(\sigma_{x_0}^2 \sin^2\beta - \sigma_S^2\right)\exp(-K_M t) + \sigma_P^2\right]. \tag{17}$$

$$[L(\beta, t)]^2 = \frac{2}{3}\left[\sigma_{x_0}^2 \exp(-K_M t) - \left(\sigma_S^2 \exp(-K_M t) - \sigma_P^2\right)\Big/\sin^2\beta\right], \ \beta_{1_0} \le \beta \le \pi/2. \tag{18}$$

Distance H_{T_N} is determined by Eq. (17) at $t = t_1$.

The presence of K_M in (17) makes it possible to describe the displacement of planes even in a load-free state. Indeed, since the distance $H_{\max}$ appearing in the definition of K_M (14) is non-zero ($H_{\max} = \sigma_{x_0} \ne 0$), formula (17) governs the displacement of planes during the annealing as $\sigma_x = 0$. Another important feature of K_M is that the intensity of the displacements depends, via σ_{x_0}, on the level of plastic pre-strain.

The behavior of function $L(\beta)$ at $t = t_1$ depends on the relationships between the values of $\sigma_S^2 \exp(-K_M t_1)$ and σ_P^2. If $\sigma_S^2 \exp(-K_M t_1) > \sigma_P^2$, $L(\beta)$ grows with increasing β, $L'(\beta) > 0$. Otherwise, $L(\beta)$ is a decreasing function of β.

The fact that $L'(\beta) > 0$ means that, during the annealing, the planes from the directions $\beta_{1_0} < \beta \le \pi/2$ travel such distances toward the origin of coordinates that their points of intersection with S_1-axis lie to the right to the cone tip.

Therefore, it can be concluded that the loading surface at the end of annealing ($t = t_1$) retains the shape formed at $t = 0$.

If $\sigma_S^2 \exp(-K_M t_1) = \sigma_P^2$, $L(\beta)$=*const*, i.e. the loading surface has a form of cone on whose tip there are all the planes that were reached by stress vector at plastic loading. The decreasing dependence of $L(\beta)$ on β means that the loading surface, being the inner envelope of tangent planes, loses the angular point (Figure 4c).

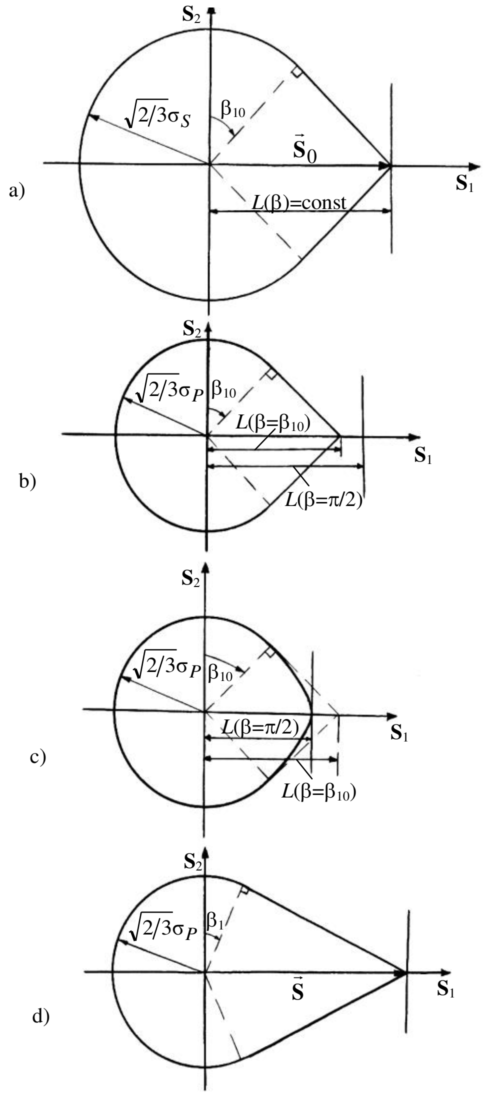


Figure 4

Evolution of loading surface in the course of preliminary MTT and subsequent creep (boundary planes are shown only)

3.3 Steady-State Creep of Metal Treated by MTT

The loading surface in creep ($\sigma_x(t) = const$), similarly to a plastic deformation, has a shape of a cone (Figure 4d) which does not vary in time. Insert H_{T_N} into (6):

$$\psi_{M_N} = \frac{2}{3}\begin{cases} (\sigma_x \sin\beta\cos\lambda)^2 - \left[(\sigma_{x_0}\sin\beta\cos\lambda)^2 - \sigma_S^2\right]\exp(-K_M t_1) - \sigma_P^2, & \Omega_{1_0}(\beta,\lambda) \\ (\sigma_x \sin\beta\cos\lambda)^2 - \sigma_P^2 & \Omega_1(\beta,\lambda) \end{cases}, \tag{19}$$

where $\Omega_{1_0}(\beta,\lambda)$ is the range of angles β and λ: $\beta_{1_0} < \beta \le \pi/2$, $0 < \lambda < \lambda_{1_0}$; $\Omega_1(\beta,\lambda)$ stands for $\beta_1 \le \beta \le \beta_{1_0}$, $\lambda_{1_0} \le \lambda \le \lambda_1$. Here we restrict ourselves to the case when the creep stress vector $\vec{\mathbf{S}}$ is such that the range $\Omega_1(\beta,\lambda)$ is greater than $\Omega_{1_0}(\beta,\lambda)$. As such, the values of λ_1 and β_1 are calculated by equating the ψ_{M_N} and λ_1 from $\Omega_1(\beta,\lambda)$ to zero [8, 11]:

$$\cos\lambda_1(\beta) = \sigma_P / [\sigma_x \sin\beta], \quad \sin\beta_1 = \sigma_P/\sigma_x . \tag{20}$$

It must be noted that the loading surface in creep without preliminary MTT is formed by the same set of tangent planes, $\Omega_1(\beta,\lambda)$. However, the planes from the diapason $\Omega_{1_0}(\beta,\lambda)$ travel greater distances on the endpoint of the $\vec{\mathbf{S}}$ than those for the treated material.

The creep strain rate after MMT, $\dot{e}_{1M}$, is determined by formulae (20), (19), (7), and (2):

$$\dot{e}_{1M} = \frac{4\pi\tilde{K}}{3r}\left\{\int\limits_0^{\lambda_1}\int\limits_{\beta_1}^{\pi/2}\left\{(\sigma_x\sin\beta\cos\lambda)^2 - \sigma_P^2\right\}\sin\beta\cos\beta\cos\lambda d\lambda d\beta - \right.$$
$$\left. -\exp(-K_M t_1)\int\limits_0^{\lambda_{1_0}}\int\limits_{\beta_{1_0}}^{\pi/2}\left\{(\sigma_{x_0}\sin\beta\cos\lambda)^2 - \sigma_S^2\right\}\sin\beta\cos\beta\cos\lambda d\lambda d\beta\right\}. \tag{21}$$

According to (5) and (2), the first integral in (21) gives the creep rate without MTT, $\dot{e}_1 = a\Phi(\sin(\beta_1))$, $a = \pi\sigma_P^2/(9r)$, and the second one does the preliminary plastic strain e_{1_0} (formulae (12) and (13)). Therefore,

$$\dot{e}_{1M} = \dot{e}_1 - \tilde{K}\exp\left[-K_M\left(e_{1_0}\right)t_1\right]\cdot e_{1_0}, \tag{22}$$

where the $\tilde{K} = const$ is obtained from Eq. (11) at $H_{\max} = \sqrt{2/3}\sigma_x = const$.

For the case when MTT is not carried out, we have $e_{1_0} = 0$ and $\dot{e}_{1M} = \dot{e}_1$. With an increase of pre-strain, the term $\tilde{K}\exp\left[-K_M\left(e_{1_0}\right)t_1\right]\cdot e_{1_0}$ first increases and then tends to zero.

On the basis of formula (22), in Figure 1 and 2, the analytical and experimental [2, 4] curves $\dot{e}_{1M} = f\left(e_{1_0}\right)$ are plotted. Good agreement between the analytical and experimental data enables us to use formula (22) to predict the steady-state creep of metals as a function of preliminary plastic deformation in the course of MTT.

Conclusions

The analysis of the evolution of loading surface in the course of mechanical-thermal treatment and subsequent steady-state creep has been studied. Depending on the value of plastic pre-strain, in the course of MTT, the loading surface can assume a conic or rounded shape. The analytical results give good agreements with experimental data.

Acknowledgement

The author expresses thanks to Prof. K. Rusinko (Budapest University of Technology and Economics, Hungary) for many useful conversations on the topics presented in this article.

References

[1] Cadek, J. (1988) *Creep in metallic materials*, Elsevier

[2] Bazelyuk, G., Kozyrskii, G., Petrunin, G., Polotskii, I. (1970) Influence of Preliminary Ultrasonic Irradiation on the High-Temperature Creep and Microhardness of Copper, *Fiz. Metal. Metalloved.,* **29**: 508-511 (in Russian)

[3] Devenyi, L., Biro, T. (2003) Investigation of Creep Cavities by Scanning Electron Microscope, *Materials Science Forum*, **414-415**: 183-187

[4] Kozyrskij, G., Kononenko, V. (1966) The Study of the Creep of Prestrained Nickel Alloy, *Fizika metallov i metallovedenije*, **22**: 108-111 (in Russian)

[5] Reger, M., Vero, B., Felde, I., Kardos, I. (2010) The Effect of Heat Treatment on the Stability of Centerline Segregation, *Strojniski vestnik – Journal of mechanical engineering*, **56**: 143-149

[6] Reger, M., Vero, B., Csepeli, Zs., Pan, J. (2004) Modeling of Intercritical Heat Treatment of DP and TRIP Steels, *Transaction of Materials and Heat Treatment*, 25: 710-715

[7] Réger, M., Louhenkilpi, S. (2003) Characterizing of the Inner Structure of Continuously Cast Sections by Using of Heat Transfer Model, *Materials Science Forum* **414-5**: 461-469

[8] Rusinko, A., Rusinko, K. (2011) *Plasticity and Creep of Metals*, Springer

[9] Rusinko, A. (2002) Analytic Dependence of the Rate of Stationary Creep of Metals on the Level of Plastic Prestrain, *J. Strength of Metals* **34**: 381-389

[10] Ruszinko, E. (2009) The Influence of Preliminary Mechanical-Thermal Treatment on the Plastic and Creep Deformation of Turbine Disks, *MECCANICA* **44**: 13-25

[11] Rusinko, A. (2010) Non-Classical Problems of Irreversible Deformation in Terms of the Synthetic Theory, *Acta Polytechnica Hungarica* **7**: 25-62

[12] Totten, G. E. (2006) *Steel Heat Treatment: Metallurgy and Technologies*, Taylor & Francis

UWB Radar Signal Processing for Positioning of Persons Changing Their Motion Activity

Jana Rovňáková, Dušan Kocur

Department of Electronics and Multimadia Communications,
Faculty of Electrical Engineering and Informatics, Technical University of Košice,
Park Komenského 13, 041 20 Košice, Slovak Republic
jana.rovnakova@tuke.sk, dusan.kocur@tuke.sk

Abstract: In many applications of ultra-wideband (UWB) radar a character of target motion is a priori not known and naturally can differ within a group of multiple targets. Usually a signal processing aimed at the positioning of only moving persons or only static persons leads to a loss of information. To solve this task, the utilization of combined processing based on detection of non-stationary signal components in the time domain and human respiratory motions in the frequency domain is proposed in this paper. The results of such processing are more reliable and robust with respect to motion activity of all monitored persons, which is demonstrated by processing of measured UWB radar signals.

Keywords: Moving targets; positioning; signal processing; static targets; UWB radar

1 Introduction

The positioning of persons by an ultra-wideband (UWB) radar has a vast number of practical applications. The examples include the tracking of people in dangerous environments (for the purposes of fire fighters and/or policemen), through rubble localization following an emergency (e.g. an explosion or earthquake) or interior monitoring (for unauthorized intruders, or for aged people, helping to ensure their health and safety). In such applications, the UWB radars have advantages over other systems due to their high spatial resolution, the usage of harmless radio waves and the ability of their stimulation signals to penetrate through different materials or obstacles [28].

Radar signal processing for the purpose of person positioning differs with respect to the target motion activity. In the case of living human beings, we can basically distinguish between moving persons whose limbs (legs, hands, head, trunk) are in motion and static persons whose limbs are motionless, but inner organs (lung, heart) still cause geometric alterations of the human body shape discernible by a high resolution UWB radar. The positioning of moving targets is usually based on monitoring of non-stationary signal components in the time domain. The target positions are then calculated analytically by localization techniques [16], [5], [12], or targets are seen as radar blobs in gradually generated radar images (radar imaging techniques, [14], [8], [6]). Commonly, the positioning is followed by a complex tracking system enabling the monitoring of changing target positions during observation time [2], [26], [17]. In the positioning of static persons, target detection is very challenging by itself. It is based on the periodical nature of the breathing or heartbeat that makes it possible to distinguish it from noise and clutter components. In the literature, most of the radar signal processing techniques are aimed at searching for the body variations caused by respiratory motions [15], [22], [11], but few of them are oriented also to cardiac-induced radar signatures [27], [3].

In our previous works [18], [9], a complete signal processing procedure for through wall tracking of multiple moving targets was introduced and tested on radar data acquired by a pseudo-noise UWB radar equipped with one transmitting and two receiving antennas [30]. The processing results have proved the ability to track a single person moving in a complex environment, such as inside fully furnished rooms, behind thick walls with high relative permittivity or even behind more walls. Successful tracking of multiple moving targets was achieved in the cases when persons did not shadow each other [10]. If during measurement an effect of mutual shadowing occurred, only a person moving nearby the radar antennas could be tracked (an effective solution is e.g. in application of UWB sensor network [19]). Experimental measurements with static persons were also realized. By using the same procedure, a motionless person breathing behind a wall could be revealed as well, but only when nobody else was moving inside the monitored area. A decreased reliability of processing results was also obtained for persons changing their motion activity, e.g. when a walking person stops for a longer time and becomes a part of the background.

As in many rescue, surveillance or security operations a character of target motion is a priori not known and naturally can differ within a group of multiple targets, an application of the procedure for positioning of only moving persons or only static persons can lead to a loss of information. Therefore, we suggest the utilization of a proper combination of both procedures. This idea is simple, but to our best knowledge its results have not yet been presented in the UWB radar literature. For that purpose, the description of the combined signal processing procedure for positioning of persons with unknown or changing motion activity represents the core of this paper. It is introduced in Section 2. The performance of the proposed procedure is demonstrated by the processing of UWB radar signals acquired for the typical scenario of moving and static target mix. It is outlined and analyzed in Section 3. Finally, some advantages and drawbacks of the introduced radar signal processing are discussed in the conclusions.

2 UWB Radar Signal Processing

In what follows, the effective procedures for the positioning of moving persons and static persons by means of UWB radar will be firstly separately described. Within them, the significance of the particular signal processing phases together with the specific methods providing stable, good and robust performance for the considered application will be outlined. Finally, a fusion principle of both procedures resulting in a combined signal processing procedure for the positioning of persons with unknown or changing motion activity will be introduced.

The radar signal processing described in this section was originally designed for signals provided by the pseudo-noise UWB radar system using the maximum-length-binary-sequence (M-sequence) as the stimulus signal [23]. As the signals acquired by the M-sequence UWB radar have a form of the impulse responses of the environment through which the stimulus signals are propagating, the same processing procedures can also be directly applied for signals obtained by means of some other kinds of UWB radars, e.g. impulse UWB radars.

2.1 Procedure for Positioning of Moving Persons

The chosen signal processing procedure [18] consists of phases responsible for the elimination of stationary clutter (methods of background subtraction), the decision about the target presence or absence (methods of detection), the estimation and association of distances from the same target (methods of time-of-arrival (TOA) estimation), the estimation of target positions (methods of localization) and finally the monitoring of target motion over time (methods of tracking). If target detection and tracking by UWB radar is realized through the walls with known parameters (thickness and relative permittivity of the wall), the phase of wall effect compensation is added between the phases of TOA estimation and target localization. The detailed description of the particular phases of radar signal processing together with the corresponding mathematical formulas is provided in [18]. The phase significance and recommended methods of signal processing are summarized below.

2.1.1 Background Subtraction

Raw radar signals can be interpreted as a set of impulse responses of surrounding through which the signals emitted by the radar were propagated. The first task of radar signal processing is to improve the signal to noise ratio. This is done by background subtraction, which especially rejects the stationary and correlated clutter, such as antenna coupling, impedance mismatch response and ambient static clutter, and allows the response of moving targets to be detected. Exponential averaging was chosen from a variety of background subtraction methods because of its robust performance and low complexity [29].

2.1.2 Target Detection

Detection represents a class of methods that on the basis of a statistical decision theory determine whether a target is absent or present in the examined radar signals. Between detectors able to provide good and robust results in the case of multi-target through wall detection by UWB radar, a constant false alarm rate (CFAR) detector can be assigned. It is based on the Neymann-Person optimum criterion providing the maximum probability of detection for a given false alarm rate. In the considered radar signal processing, the CFAR detector that assumes a Gaussian clutter model has been applied [4].

2.1.3 TOA Estimation

Binary data representing the detector output form a noticeable trace of the moving targets. It represents the time of arrival (TOA) of the electromagnetic waves reflected by the target for the particular instants of the observation time. As the range resolution of UWB radars is considerably high with regard to the physical dimensions of the targets to be detected, the targets are usually represented by more TOA values in the detector output. In order to simplify the target localization, such distributed targets are replaced by simple targets; i.e. the target position in every observation time instant is given by only one TOA. This phase of radar signal processing is referred to as the TOA estimation. For its realization, a novel algorithm entitled TOA association has been proposed in [20]. It further enables the combining of the TOAs estimated from both receiving antennas into couples from which the positions of the potential true targets can be computed during localization phase. This part of algorithm represents a data-association phase and is responsible for the de-ghosting task solution.

2.1.4 Wall Effect Compensation

The propagation of electromagnetic waves through a wall results in a delay time of signals reflected by targets moving behind the wall. This means that the TOA estimated by the previous phase of radar signal processing are time shifted because of the wall presence. Their correction can be achieved by the subtraction of the mentioned delay time, whereas its estimation is the task of the wall effect compensation phase. The method referred to as the target trace correction of the 2nd kind [21] provides promising results in this area. To use this method, the wall parameters, such as the permittivity, permeability and thickness of the wall, must be known in advance, or they can be estimated very effectively with the same M-sequence UWB radar by using the method described in [1].

2.1.5 Target Localization

The aim of the localization phase is to determine the target coordinates in the defined coordinate systems whereby the target locations estimated in consecutive time instants create the target trajectory. As the input of this radar signal processing phase, the estimated and corrected TOA couples are used. Because the considered UWB radar system consists of one transmitting and two receiving antennas, only the non-iterative direct method of localization can be employed. In this case, the target coordinates are simply calculated by trilateration methods as intersections of two ellipses formed on the basis of the estimated TOA couples and known coordinates of the transmitting and receiving antennas [16].

2.1.6 Target Tracking

The particular locations of the targets are estimated with certain random error usually described by its probability distribution function. Taking into account this model of target position estimation, the target trajectory can be further processed via tracking algorithms. They provide a new estimation of the target location based on the previous positions of the target. Usually, the tracking results in a decrease in the target trajectory error and includes trajectory smoothing. In the case of multiple targets, track filtering must also deal with track maintenance and with the problem of determining which measurements to associate with which targets being tracked. From a wide spectrum of tracking algorithms, the multiple target tracking (MTT) system using a linear Kalman filtering has been chosen as the method to enclose the complex procedure of the UWB radar signal processing applied for through wall tracking of the multiple moving targets [7].

2.2 Procedure for Positioning of Static Persons

The basis of the signal processing procedure for static persons positioning was originally introduced in [24]. It is quite close to the procedure described in Section 2.1 except that body movements are markedly restricted. Therefore, the signal to clutter ratio will further decrease and the only feature to distinguish the persons from static objects is their small movement due to breathing. There are few features that serve as a basis for methods to enhance the response from a breathing person:

- Human breathing can be considered as periodical motion over a certain interval of time. The frequency of breathing can change slowly with time, but it should always be within a frequency band of about 0.2-0.5 Hz.
- The geometrical variations of the thorax caused by breathing are usually quite a bit less than the range resolution of the radar.

- The echo due to a breathing person is extremely weak, all the more so when the static person is behind an obstacle (e.g. rubble, wall, snow) that strongly attenuates the sounding waves.
- The distance from antennas to the breathing persons does not change during the measurement.

Taking into account these features, the phase of background subtraction is supplemented by a breathing enhancement and followed by transformation of the radar signals into the frequency domain. Consequently, the estimation of power spectral density is used for the breathing detection task. The remaining processing phases, i.e. the TOA estimation, wall effect compensation and target localization, are the same as in the procedure for the positioning of moving persons, but they are applied now on an essentially reduced data set, the size of which corresponds to the number of detected static persons.

It is useful to mention that the UWB radar device applied for the positioning of static persons has a great influence on the success of breathing detection. It can be explained as follows. As the detection of small movements is based on observing the variations of steep signal flanks, the noise should be as small as possible there. It is known that the jitter provokes additional noise on the signal flanks and, therefore, the short time stability of the radar device is of major importance for such applications (the considered M-sequence UWB radar behaves excellently with respect to this point) [25].

2.2.1 Background Subtraction and Breathing Enhancement

The method of exponential averaging recommended for background subtraction in Section 2.1.1 can be used to advantage also in this case [29]. However, the weighing factor controlling the amount of averaging in the background estimation should be set now in such a way as to smooth out high frequency variations and reveal long term trends in the background estimation (i.e. it provides low-pass narrow band filtering). It is done by choosing a longer fraction of the previous estimate of impulse response with subtracted background and a smaller fraction of the actual measured impulse response.

In order to further improve the signal-to-noise ratio of a static target echo, the impulse response with subtracted background is applied to a so-called range filter before the target detection [13]. The range filter helps to improve the signal-to-noise by reducing the clutter residue and noise resulting from the de-correlation of any radio frequency interference due to pseudo-random code transmitted by the radar.

Additionally, breathing as a narrow band process can also be enhanced by the use of low-pass filtering. Here, a low-pass filtering with a cut-off frequency higher than the highest frequency of breathing (e.g. higher than 1 Hz) can be applied along the observation time axis for each propagation time instant to suppress high-frequency noise [24].

2.2.2 Estimation of Power Spectral Density

To extract the breathing rate, a horizontal Fast Fourier Transform (FFT, along the time observation axis for each propagation time instant) is applied on the radar signals after background subtraction and breathing enhancement. The frequency of breathing can change with the observation time. However, the bandwidth of breathing from one person under observation with radar is likely to be considerably less than a priory bandwidth as determined for the whole range of respiratory activity for all individuals. Thus, the total energy contained within the frequency window can serve as an indicator as to whether breathing is present. Finally, the FFT-based Welch periodogram is used for estimating the power spectral density (PSD) of the radar signals in the direction of the observation time [24].

2.2.3 Target Detection

For the detection of static persons, the CFAR detector mentioned in Section 2.1.2 or a simpler threshold detector can be advantageously used [4]. The detector is applied on the data set represented by the estimated PSD. If a breathing person is present in the monitored area, the detector binary output should gain values “1” between frequencies 0.2-0.5 Hz corresponding to the expected breathing rate of human being. If more static persons are situated inside the area, values “1” should occur in more propagation time instants. In order to later estimate for every detected target only one spatial position, the frequency responses from interval 0.2-0.5 Hz are simply summed to one frequency response. Such a response represents then the input to the TOA estimation phase.

2.2.4 TOA Estimation

The TOA estimation phase is realized via the TOA association method (Section 2.1.3) [20]. The estimated TOA for every detected static target represents a round trip time between the transmitting antenna – the target – and the receiving antenna. The TOA multiplied by the light propagation velocity gives the distance between them.

2.2.5 Wall Effect Compensation

Instead of the target trace correction of the 2nd kind outlined in Section 2.1.4, the simpler target trace correction of the 1st kind can be used. The approximate delay time is calculated from the elementary equation given in [21].

2.2.6 Target Localization

If both receiving antennas detect the presence of a breathing person, its position will be calculated as the intersection of the ellipses given by the couple of the TOAs associated during the TOA estimation phase and corrected during the wall effect compensation phase (as long as the wall parameters are known) [16]. If only one receiving antenna confirms a target presence, at least the incomplete positioning of the static person based on the distance from this antenna can be used. The possible locations of the person are then given by the half-ellipse situated inside a monitored area.

2.3 Combined Procedure for Positioning of Persons with Unknown or Changing Motion Activity

The combined procedure exploits the parallel processing of measured radar signals via the procedure for the positioning of moving persons (PPMP) and the procedure for the positioning of static persons (PPSP) described in the previous sections. The procedure flowchart is depicted in Figure 1. It can be seen that PPMP runs consecutively as soon as the impulse responses (IRs) are acquired from the two receiving channels. This results from the fact that the low complex methods of PPMP are always applied only on the actual couple of the IRs (IR from receiving antenna Rx1, IR from receiving antenna Rx2). On the other hand, PPSP requires for proper functioning a larger set of IRs to be able to distinguish periodical breathing from noise and clutter components.

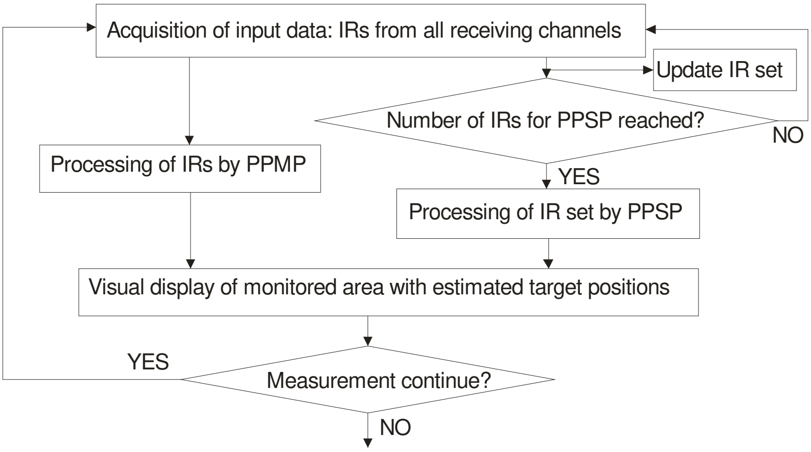

Acquisition of input data: IRs from all receiving channels
Update IR set
Number of IRs for PPSP reached?
NO
Processing of IRs by PPMP
YES
Processing of IR set by PPSP
Visual display of monitored area with estimated target positions
YES
Measurement continue?
NO

Figure 1
The flowchart of the combined procedure for positioning of persons with unknown or changing motion activity

The processing of the IR set in the frequency domain is also more time consuming. Therefore, the information about static person positions comes in at slower time intervals. Subsequently, the outputs of PPSP are depicted on the same visual display of the monitored area as the PPMP outputs (Figure 1). Although the information about static person positions are delayed compared to the true target positions, they make the radar processing results more reliable and robust with respect to the motion activity of all monitored persons.

3 Experimental Results

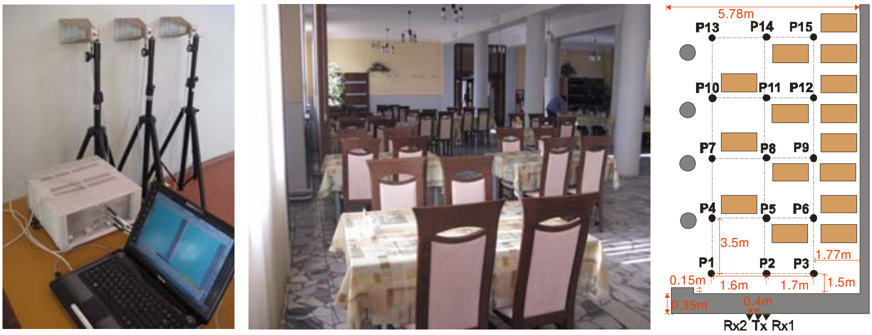
The performance of the proposed combined PPMP and PPSP procedure is demonstrated by the processing of UWB radar signals acquired for a measurement scenario with one static person and one moving person changing his motion activity during measurement. In the next sections, the applied UWB radar device, chosen scenario and obtained processing outputs are described in detail.

3.1 UWB Radar Device

An experimental version of the M-sequence UWB radar system equipped with one transmitting and two receiving horn antennas was utilized for the measurement (Figure 2(a)) [23]. The system clock frequency of the radar device is about 4.5 GHz, which results in an operational bandwidth of about DC-2.25 GHz. The M-sequence order emitted by the radar is 9; i.e. the impulse response covers 511 samples regularly spread over 114 ns. This corresponds to an observation window of 114 ns leading to an unambiguous range of about 17 m. The measurement speed reaches approximately 13.5 impulse responses per second.

3.2 Measurement Scenario

A part of a school dining room of a size approximately 6 m x 17 m was chosen for the monitored area. As can be seen from photo in Figure 2(b), the room was furnished with high wooden chairs and tables with metallic legs. Within the analyzed scenario, two persons, labeled as target A and target B, were present inside the monitored area. Target A was at first walking around tables according a rectangular trajectory given by the reference positions P1-P13-P15-P3-P1 depicted in the measurement scheme in Figure 2(c). Afterwards, the person remained at position P1 until the end of measurement.



(a) (b) (c)

Figure 2

Measurement setup: (a) M-sequence UWB radar located behind the wall, (b) interior of the monitored area, (c) scenario scheme with the positions of radar antennas and the reference positions

Target B was sitting motionless during whole measurement (2 min 36 s) at position P5.

The M-sequence UWB radar system (Figure 2(a)) was located behind a 0.35 m-thick brick wall, depicted in Figure 2(b). All antennas were placed along a line with the transmitting antenna Tx in the middle of Rx1 and Rx2, with the distances between adjacent antennas set at 0.4 m (Figure 2(c)). There was no separation between radar antennas and the wall (Figure 2(a)).

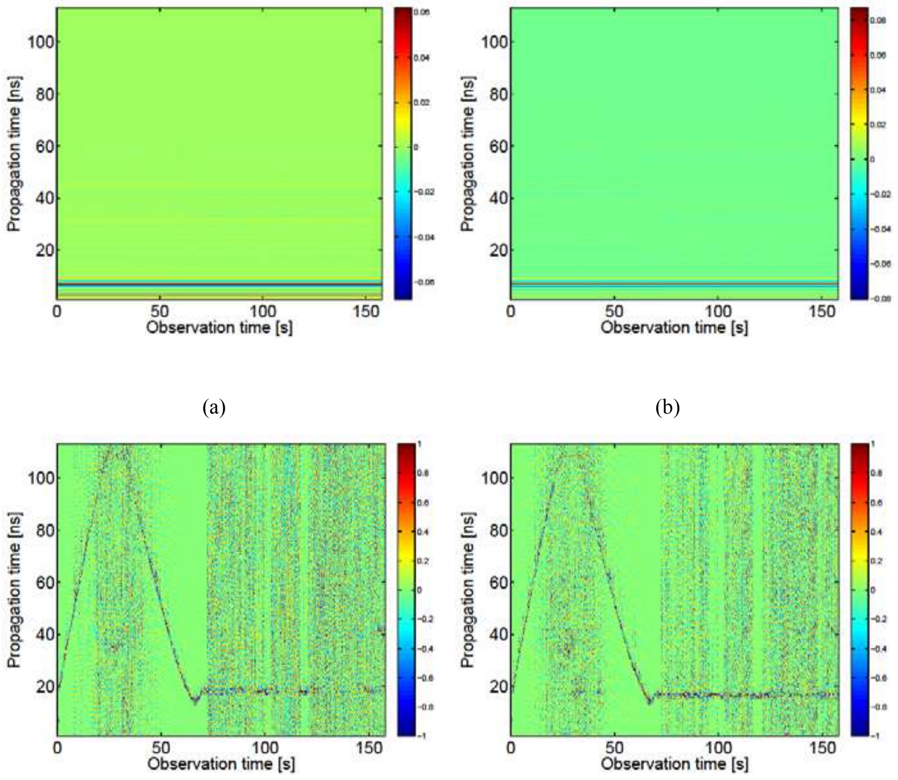
3.3 Signal Processing Outputs

As was explained in Section 2.3, the combined procedure exploits parallel processing of the measured radar signals by PPMP and PPSP. Therefore, in what follows, the outputs from every processing phase are shown at first for PPMP, then subsequently for PPSP, and finally for the combined procedure.

The raw radar signals corresponding to the measurement scenario and obtained by the receiving channels Rx1 are Rx2 are depicted in Figure 3(a) and 3(b), respectively. They have the form of a radargram, i.e. a two dimensional picture in which the vertical axis is related to the propagation time of the impulse response and the horizontal axis is related to the observation time. In these radargrams, only a cross-talk signal and the reflections of the emitted electromagnetic wave from the wall can be viewed, as they are very strong in comparison with the weak signals scattered by the persons.

This situation is changed after the phase of background subtraction, when the primary trace of target A has arisen in the radargrams (Figure 3(c)-3(d)). The trace shape corresponds with the target motion; i.e. at first, the person is retreating from the radar antennas, then approaching them and finally remaining at approximately the same distance from the radar. Target A was not totally motionless while standing (movements with hands and head, turning about) and therefore the primary target trace is partially noticeable also in the second half of radargrams depicted in Figure 3(c)-3(d). The presence of target B is not observable from these figures because this was evaluated as a static background and subtracted from the data.

The CFAR detector outputs are shown in Figure 3(e)-3(f). The false alarm rate was adjusted to higher values in order to detect also targets from noisy signals. This results in a greater probability of target detection but also in a higher number of the false alarms (randomly distributed points in Figure 3(e)-3(f)). By comparing Figure 3(e) and Figure 3(f), it can be seen that channel Rx1 has received weaker reflections from the standing person. This is caused by the fact that the standing person position (the reference position P1) was located nearer to antenna Rx2.

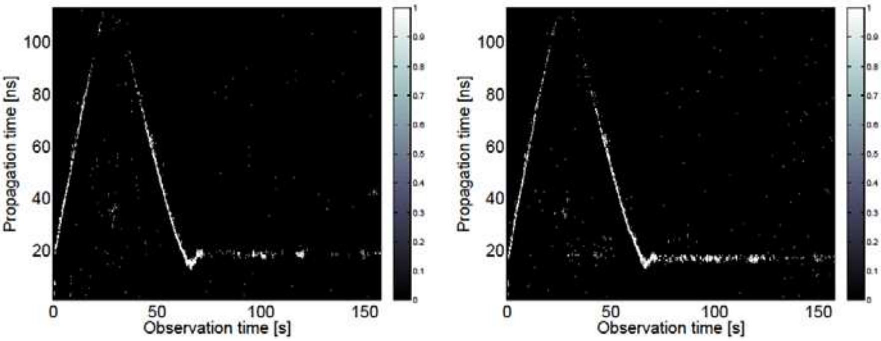


(a)

(b)

(c)

(d)



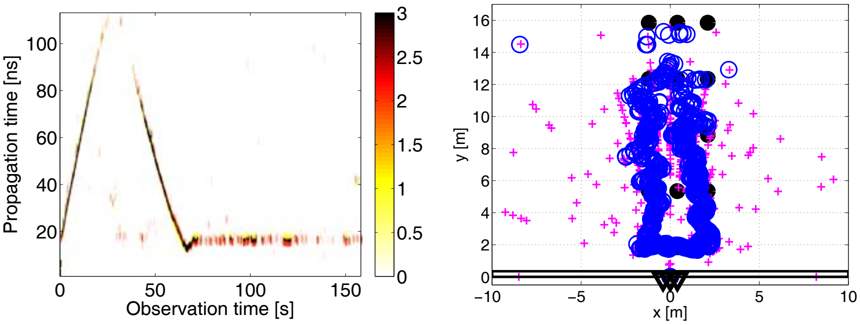
(e) (f)

Figure 3
Signal processing results from PPMP: Part I. Radargram depictive raw radar signals: (a) channel Rx1, (b) channel Rx2; Radargram depictive signals with subtracted background: (c) channel Rx1, (d) channel Rx2; Radargram depictive detector outputs: (e) channel Rx1, (f) channel Rx2

A similarity of radargrams obtained by both receiving channels can be seen in Figure 4(a), where the TOA couples belonging to the same target are highlighted in black. This similarity results from the symmetric and small distance between the antennas. In Figure 4(a), the TOA estimated from channels Rx1 and Rx2 are outlined in yellow and red, respectively, and both are artificially widened for better visibility. Their conjunction implies that both receiving channels captured the relevant reflections from the same target. Not associated TOA are considered to be false alarms. In this way ghost generation is avoided.

As the relative permittivity of the wall was not measured for this scenario, the wall effect compensation phase has been omitted from the processing. The outputs of the localization phase are depicted in Figure 4(b) with magenta crosses. They were computed based on the TOA couples from Figure 4(a). Because at further distances the reflections of the moving target were captured only alternately by the receiving channels, the target locations could not be estimated in these observation time instants (around 20-40 ns in Figure 4(a)). Similarly, the positions of the standing person were computed in only a few observation time instants, when the person moved his hands or turned about.

The final result from PPMP is illustrated in Figure 4(b) with circles. The circles create a target track estimated by the MTT system. In comparison with the target positions estimated in the localization phase ("+"), the track is more smoothed and complemented about missing positions, if it was possible. The estimated track corresponds well with the true rectangular trajectory of moving target A except for at the furthest areas, in which the reflection from the person was very weak. While standing at the same position, target A was detected too, but only in a few observation time instants (left down corner of the rectangular track in Figure 4(b)). The sitting target B was not detected at all by PPMP.



(a) (b)

Figure 4

Signal processing results from PPMP: Part II. (a) Radargram depictive artificially widened TOA estimated from Rx1 (values 1) and Rx2 (values 2) , values 3 indicate associated TOA couples; (b) Monitored area depictive the positions of target A estimated in the localization phase (crosses) and in the tracking phase (circles), full circles – reference positions, triangles – radar antennas, lines - wall

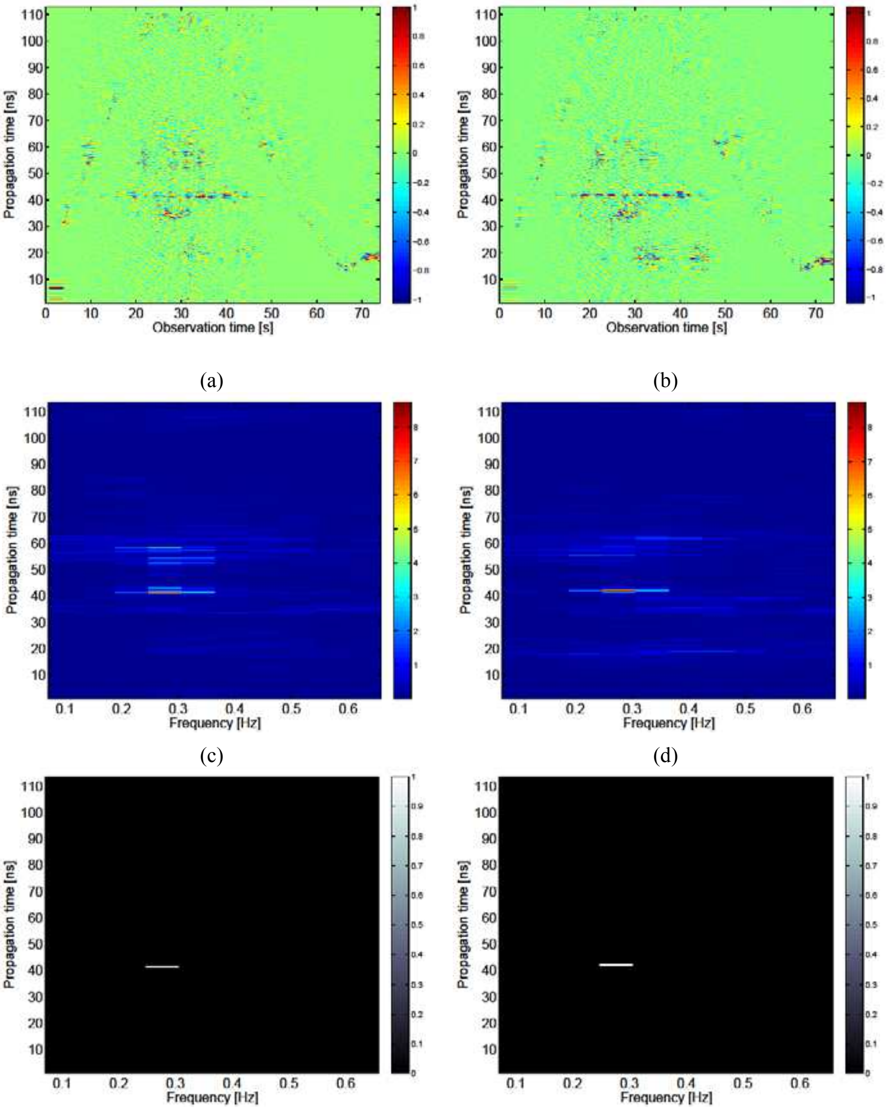
As was mentioned in Section 2.3, to function properly, PPSP requires a larger set of IRs in order to be able to distinguish periodical breathing from noise and clutter components. The size of the IR set for PPSP processing was set to approximately 1000 IRs on the basis of experimental testing. As the measurement rate of the used M-sequence UWB radar is around 13.5 IRs per second, such amount was reached roughly after 75 s. The signal processing outputs of PPSP from the first set of IR, i.e. from the observation time interval 1 s - 75 s, are shown in Figure 5, and from the second set of IR, i.e. from the observation time interval 75 s - 150 s, are shown in Figure 6.

The raw radar signals are at first processed by the phase of background subtraction and breathing enhancement (Figure 5(a)-(b)). In the obtained radargrams, the components pertaining to the moving target A are suppressed, and in contrast, the components pertaining to the motionless target B are enhanced. The primary trace of target B has the shape of a discontinuous line occurring around the propagation time 40 ns.

The estimation of PSD for both channels is depicted in Figure 5(c)-(d). It serves for the extracting of the breathing rate. As can be observed from these figures, a few values are highlighted in the frequency window 0.2-0.5 Hz, which corresponds to the typical breathing rate of human beings.

The application of the threshold detector correctly designates a presence of one static person (Figure 5(e)-(f)). His breathing rate reached values of around 0.25-0.3 Hz.

The signal processing outputs from the phase of TOA estimation and localization are given in Figures 5(g) and (h), respectively. The summation of the frequency responses from interval 0.2-0.5 Hz to one frequency response allows us to estimate only one couple of the TOA (Figure 5(g)) and one target location (Figure 5(h)) for every detected person. From the first IR set, the PPSP correctly detected and localized the sitting person. As could be expected, the moving target A was not detected by PPSP.



(a)

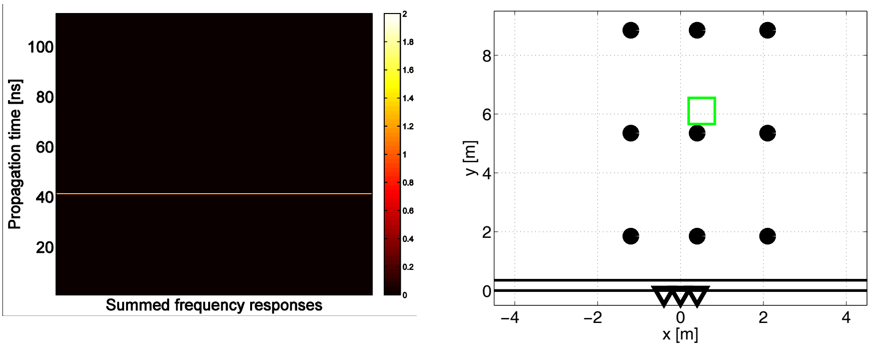
(b)

(c)

(d)

(e)

(f)

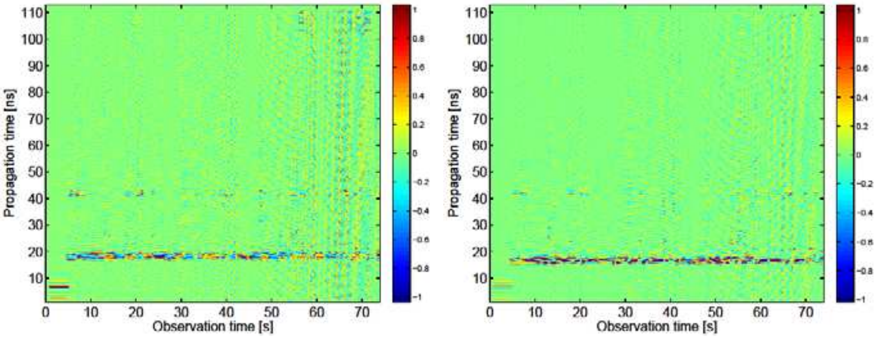


(g) (h)

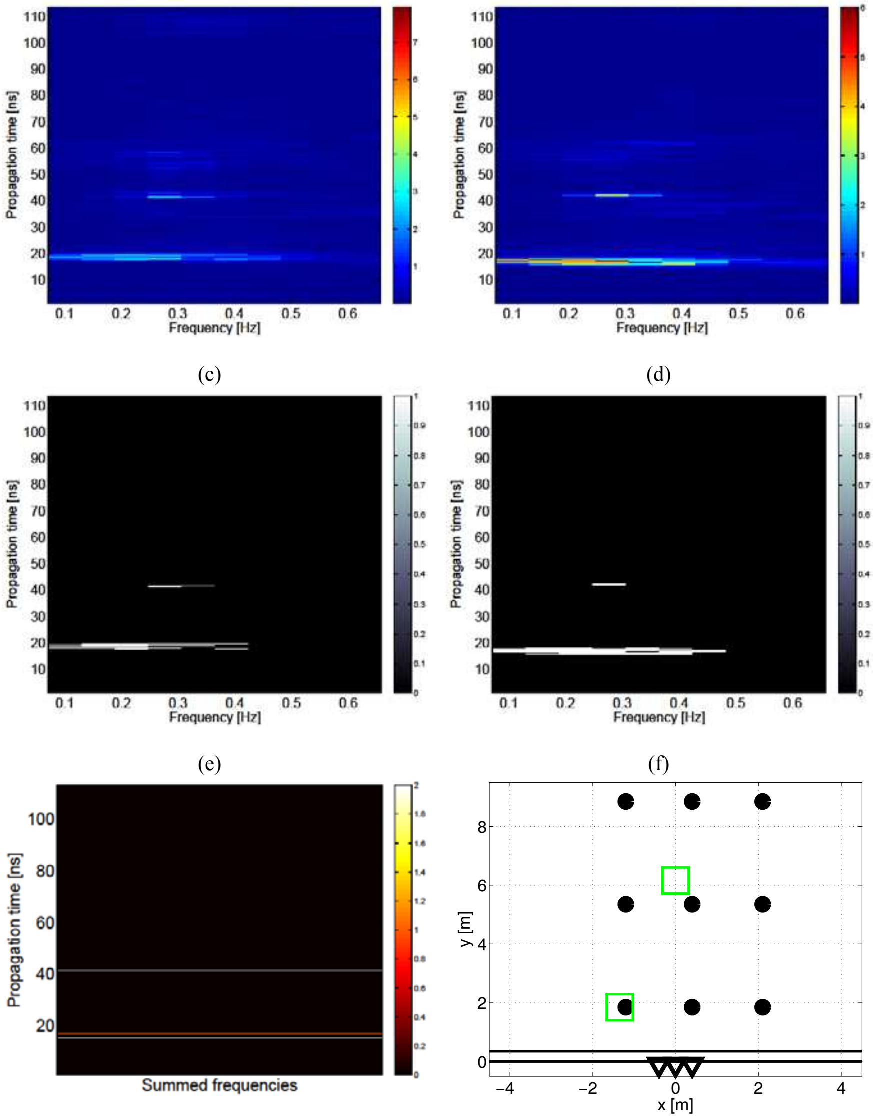
Figure 5

Signal processing results from PPSP: Part I. Signals with subtracted background and enhanced breathing: (a) channel Rx1, (b) channel Rx2; Estimated PSD: (c) channel Rx1, (d) channel Rx2; Detector output: (e) channel Rx1, (f) channel Rx2; (g) Estimated TOA couple (the values from both channels are very close each other); (h) Estimated position of a static target (square), circles – reference positions, triangles – radar antennas, lines - wall

After acquiring the second IR set of 1000 IRs from the observation time interval 75 s - 150 s, the same signal processing steps were realized. The results are depicted in Figure 6. The PPSP detected and localized the standing target A in addition to the sitting target B (Figure 6(e)-(h)). The breathing rate of target A includes more components due to the motion activity of this person (Figure 6(c)-(f)). It can be explained by the fact that it took some time to slow down breathing after walking and the person was also not totally motionless at his position. During the TOA estimation phase, two couples of TOA were associated (Figure 6(g)), which resulted in the computation of two final target locations (Figure 6(h)). As can be seen from this figure, the estimated positions correspond very well with the true target positions.



(a) (b)



(c) (d)

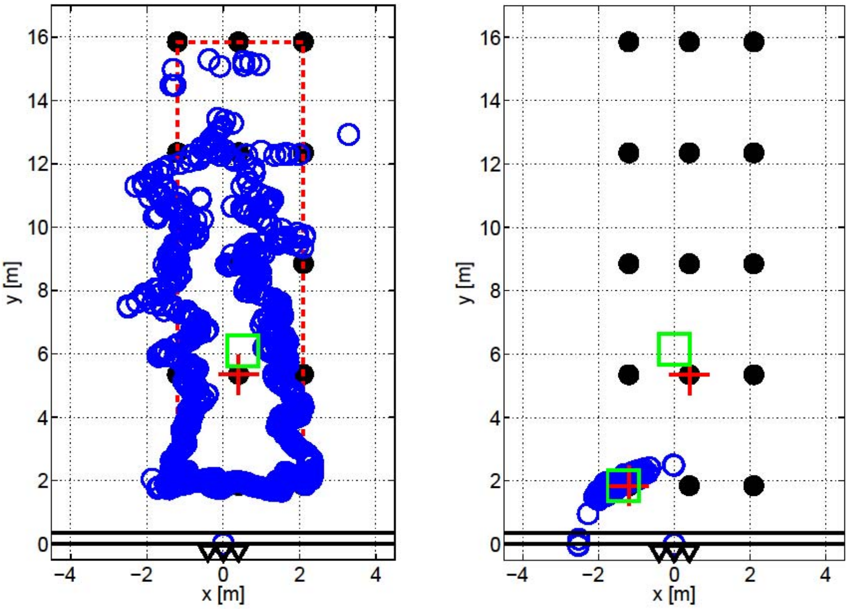
(e) (f)

(g) (h)

Figure 6

Signal processing results from PPSP: Part II. Signals with subtracted background and enhanced breathing: (a) channel Rx1, (b) channel Rx2; Estimated PSD: (c) channel Rx1, (d) channel Rx2; Detector output: (e) channel Rx1, (f) channel Rx2; (g) Estimated TOA couples; (h) Estimated positions of static targets (squares), circles – reference positions P1-P9, triangles – radar antennas, lines - wall

The final signal processing outputs from the combined procedure are obtained by fusion of results from PPMP and PPSP (Figure 7). During the first part of the measurement scenario, PPMP correctly tracks the moving person but only thanks to processing by PPSP can the motionless sitting person be detected and localized (Figure 7(a)). The imprecision between true and estimated target positions is caused by the complexity of the real environment (mainly the presence of the wall and furniture and the shadowing of target A by target B). The results for the second part of the measurement scenario are illustrated in Figure 7(b). The PPSP accurately detected and localized both static persons, but such output was available only after acquiring and processing the required IR set. Till then, PPMP was providing information about the negligible movements of standing person.



(a) (b)

Figure 7

Final signal processing results from combined procedure: obtained for (a) the first part of the measurement scenario (observation time interval 1 s - 75 s), (b) the second part of the measurement scenario (observation time interval 75 s - 156 s); Red crosses / dashed line – true target position / trajectory, blue circles – target positions estimated by PPMP, green square – target positions estimated by PPSP, black full circles – reference positions, black triangles – radar antennas, black lines - wall

Conclusions

The presented experimental results, as well as the results obtained by the processing of similar measurement scenarios, demonstrate the ability of the proposed combined processing procedure to obtain from the same measured set of UWB radar signals more extensive and more precise information about the positions of monitored persons located behind an obstacle. This advantage is reached at the expense of higher computational complexity and time consumption. The optimalization of software and hardware can reduce this disadvantage. From the point of view of software, the procedure for the positioning of static persons should be especially improved. The current version utilizes very simple methods which can be replaced by more advanced approaches known from the latest research works. From the point of view of hardware, an adequate increase in the measurement rate can be very conducive for decreasing the delay in processing the results.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under contract No. LPP-0080-09 and by the Slovak Cultural and Educational Grant Agency under contract No. 010TUKE-4/2012.

References

[1] Aftanas M., Sachs J., Drutarovsky M., Kocur D.: Efficient and Fast Method of Wall Parameter Estimation by Using UWB Radar System, Frequenz, Vol. 63, No. 11-12, Nov. 2009, pp. 231-235

[2] Blackman S. S., Popoli R.: Design and Analysis of Modern Tracking Systems. Artech House Publishers, August 1993

[3] Bugaev A. S., Chapursky V. V., Ivashov S. I., Razevig V. V., Sheyko A. P., Vasilyev I. A.: Through Wall Sensing of Human Breathing and Heart Beating by Monochromatic Radar, Proceedings of Tenth International Conference on Ground Penetrating Radar, Delft, The Netherlands, June 2004, pp. 291-294

[4] Dutta P. K., Arora A. K., Bibyk S. B.: Towards Radar-enabled Sensor Networks, Proceedings of the Fifth International Conference on Information Processing in Sensor Networks, Apr. 2006, pp. 467-474

[5] Gezici S., Tian Z., Giannakis G. B., Kobayashi H., Molisch A. F., Poor H. V., Sahinoglu Z.: Localization via Ultra-Wideband Radios: A Look at Positioning Aspects for Future Sensor Networks, IEEE Signal Processing Magazine, Vol. 22, No. 4, July 2005, pp. 70-84

[6] Cheney M., Borden B.: Imaging Moving Targets from Scattered Waves, Inverse Problems, Vol. 24, No. 3, Jun 2008

[7] Khan J., et al.: Multiple Target Tracking System Design for Driver Assistance Application," Proceedings of Conference on Design and Architectures for Signal and Image Processing (DASIP), Belgium, 2008

[8] Kocur, D., Gamec J., Švecová M., Gamcová M., Rovňáková J.: Imaging Method: An Efficient Algorithm for Moving Target Tracking by UWB Radar, Acta Polytechnica Hungarica, Vol. 7, No. 3, 2010, pp. 6-24

[9] Kocur D., Rovňáková J., Švecová M.: Through Wall Tracking of Moving Targets by M-Sequence UWB Radar, The Springer's book series "Studies in Computational Intelligence" with volume title "Computational Intelligence in Engineering", Volume 243, ISSN 1860-949X, 2009, pp. 349-364

[10] Kocur D., Rovňáková J., Urdzík D.: Short-Range UWB Radar Application: Problem of Mutual Shadowing between Targets, Elektrorevue, ISSN 1213-1539, Vol. 2, No. 4, December 2011, pp. 37-43

[11] Li J., Zeng Z., Sun J., Liu F.: Through-Wall Detection of Human Being's Movement by UWB Radar, Geoscience and Remote Sensing Letters, IEEE, Volume 9, Issue 6, 2012, pp. 1079-1083

[12] Mao G., Fidan B., Anderson B. D. O.: Wireless Sensor Network Localization Techniques, Computer Networks: The International Journal of Computer and Telecommunications Networking, Vol. 51, No. 10, July 2007, pp. 2529-2553

[13] Nag S., Barnes M.: A Moving Target Detection Filter for an Ultra-Wideband Radar, Proceedings of IEEE Radar Conference, May 2003, pp. 147-153

[14] Narayanan R.: Through Wall Radar Imaging Using UWB Noise Waveforms, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 5185-5188

[15] Nezirovic A., Yarovoy A. G., Ligthart L. P.: Signal Processing for Improved Detection of Trapped Victims Using UWB Radar, IEEE Transactions on Geoscience and Remote Sensing, Vol. 48, No. 4, 2010, pp. 2005-2014

[16] Paolini E., Giorgetti A., Chiani M., Minutolo R., Montanari M.: Localization Capability of Cooperative Anti-Intruder Radar Systems, EURASIP Journal on Advances in Signal Processing, Vol. 2008, March 2008, pp. 1-14

[17] Ristic B., Arulampalam S., Gordon N.: Beyond the Kalman Filter: Particle Filters for Tracking Applications, Artech House, 2004

[18] Rovňáková J.: Complete Signal Processing for through Wall Tracking of Moving Targets, LAP LAMBERT Academic Publishing, Germany, September 2010

[19] Rovňáková J., Kocur D.: Short Range Tracking of Moving Persons by UWB Sensor Network, Proceedings of the 8th European Radar Conference (EuRAD), Manchester, UK, Oct. 2011, pp. 321-324

[20] Rovňáková J., Kocur D.: TOA Estimation and Data Association for Through Wall Tracking of Moving Targets, EURASIP Journal on Wireless Communications and Networking, The special issue: Radar and Sonar

Sensor Networks, Volume 2010, Article ID 420767, September 2010, pp. 1-11

[21] Rovňáková J., Kocur D.: Compensation of Wall Effect for Through Wall Tracking of Moving Targets, Radioengineering: Special Issue on Workshop of the COST Action IC0803, Vol. 18, No. 2, June 2009, pp. 189-195

[22] Sachs J., Helbig M., Herrmann R., Kmec M., Schilling K., Zaikov E., Rauschenbach P.: Trapped Victim Detection by Pseudo-Noise Radar, Proceedings of the 1st International Conference on Wireless Technologies for Humanitarian Relief (ACWR), 2010, pp. 265-272

[23] Sachs J., Herrmann R., Kmec M., et al.: Recent Advances and Applications of M-Sequence based Ultra-Wideband Sensors, Proceedings of International Conference on Ultra-Wideband, Singapore, 2007

[24] Sachs J., Zaikov E., Kocur D., Rovňáková J., Švecová M.: Ultra Wideband Radio Application for Localisation of Hidden People and Detection of Unauthorised Objects. D13 Midterm report on person detection and localisation, Project RADIOTECT, COOP-CT-2006-032744, July 2008

[25] Sachs J., Aftanas M., Crabbe S., et al.: Detection and Tracking of Moving or Trapped People Hidden by Obstacles using Ultra-Wideband Pseudo-Noise Radar, Proceedings of the 5th European Radar Conference (EuRAD), Amsterdam, The Netherlands, October 2008, pp. 408-411

[26] Švecová M., Kocur D.: Taylor Series-based Tracking Algorithm for Through Wall Tracking of a Moving Persons, Acta Polytechnica Hungarica, Vol. 7, No. 1, 2010, pp. 5-21

[27] Thiel F., Hein M. A., et al.: Exploring the Benefits of Ultra-Wideband Radar for High- and Ultra-High Field Magnetic Resonance Imaging, Proceedings of Microwave Conference (EuMC), Sept. 2009, pp. 866-869

[28] Withington P., Fluhler H., Nag S.: Enhancing Homeland Security with Advanced UWB Sensors, Microwave Magazine, IEEE, Vol. 4, No. 3, Sep. 2003, pp. 51-58

[29] Zetik R., Crabbe S., Krajnak J., Peyerl P., Sachs J., Thomä R.: Detection and Localization of Persons Behind Obstacles Using M-Sequence through-The-Wall Radar, Proceedings of SPIE – Defense & Security Symposium, Vol. 6201, Orlando, Florida, USA, May 2006, pp. 62010I

[30] Zetik R., Sachs J., Thoma R. S.: UWB Short-Range Radar Sensing: The Architecture of a Baseband, Pseudo-Noise UWB Radar Sensor, IEEE Instrumentation & Measurement Magazine, Vol. 10, No. 2, April 2007, pp. 39-45

Data Envelopment Analysis of Higher Education Competitiveness Indices in Europe

József Kabók[1], Tibor Kis[2], Maria Csüllög[2], Imre Lendák[3]

[1] Provincial Secretariat for Science and Technological Development, Bulevar Mihajla Pupina 16, 21000 Novi Sad, Serbia
Jozef.Kabok@vojvodina.gov.rs

[2] University of Novi Sad, Faculty of Economics, Segedinski put 9 -11, 24000 Subotica, Serbia
tkis@ef.uns.ac.rs; cilegm@ef.uns.ac.rs

[3] University of Novi Sad, Faculty of Technical Sciences, Dositeja Obradovića 6, 21000 Novi Sad, Serbia
lendak@uns.ac.rs

Abstract: In this paper, the data envelopment analysis method is used to examine the competitiveness of higher education in selected countries / regions in Europe. The competitiveness of higher education is analyzed on the basis of available information on the chosen indicators, interacting with a model of investment which is applied in the field of higher education. The purpose of the research is to determine the level of competitiveness of higher education of the Republic of Serbia and its Autonomous Province of Vojvodina, as a European region, compared to selected European countries. The research results indicate that the application of the new investment model would improve the unsatisfactory competitiveness of the higher education of the Republic of Serbia and AP Vojvodina. Because of the large differences in the competitiveness of higher education among the analysed countries / regions, the research results point to the need for a unified approach in creating a development strategy and improving the competitiveness of higher education in Europe.

Keywords: competitiveness index; investment model; higher education

1 Introduction

Higher education is of particular importance for the economic competitiveness of any society, as higher education institutions generate knowledge and develop expertise and skills which enable individuals to achieve their personal goals as well as become valuable members of society. The competitiveness of higher education should be implemented so that it can be enjoyed by all institutions, regions and countries, and contribute to a global society based on education [25].

The choice of a higher education investment model has a decisive impact on the competitiveness of higher education. It is shown that the quality of teaching, the scope and structure of investment, the number of students as well as the competitiveness of higher education all depend on the investment model. Different investment models result from the social conditions in certain countries, the method of implementation of the Bologna Declaration [24], as well as historical, educational and cultural characteristics. The choice of the investment model depends on the budget capacity but also on the commitment of the state regarding the amount of public funds spent on investment in higher education.

Investment in higher education in Serbia is insufficient. In 2010, it amounted to €1,201.90 of budgetary resources per student [22], and in AP Vojvodina, as a European region, €1,076.60 per student [22]. Out of the 30 countries/regions analysed (26 countries of the European Union, all countries with the exception of Luxembourg, which could not be included in the research because of the inadequacy of information in the field of higher education, as well as Serbia, AP Vojvodina, Croatia and FYR Macedonia), Serbia was 28th, and the AP Vojvodina 29th in front of the FYR of Macedonia [10]. These data indicate the need to implement new investment models for higher education in Serbia, which will encourage investment in higher education, especially in its competitiveness and ability to be a participant in the competitive market of higher education of the European Union.

This paper analyses the state of the competitiveness of higher education in Serbia and AP Vojvodina, through observation and analysis of selected indicators of competitiveness and through calculating the index of competitiveness of higher education. The goal of this paper is to present an overview of the possibilities for increasing the level of the competitiveness of higher education in Serbia by implementing a new investment model.

One of the aims of the present study is an econometric analysis of the rankings of Serbia and AP Vojvodina, in terms of the competitiveness of higher education in relation to the aforementioned European countries in the period from 2006 to 2010. Using econometric analysis, this study determines changes in the rank and competitiveness index of higher education in Serbia and AP Vojvodina in relation to the aforementioned European countries, applying a new model of investment.

Since the present study analyses the competitiveness of higher education, the results are significant for both the professional and the academic community. The importance of the study is also reflected in the fact that higher education has a crucial influence on the creation of a knowledge based society, on learning and on innovation, an important political as well economic goal of all countries of the European Union in line with the Lisbon Agenda [18]. Therefore, the study is important for the strategy-makers in the European area of higher education, but also for development strategy-makers in higher education as well as institutions of higher education in Serbia and AP Vojvodina.

The paper consists of five chapters. After the introductory section, the second chapter refers to the literature review dealing with related research. The third section describes the research methodology, and the fourth analyses the results of research compared to the index of competitiveness of higher education in the selected countries / regions in Europe. The conclusion, based on the analysis of survey results, proposes measures and activities to the strategy and development policy makers of higher education in the European Union, Serbia and AP Vojvodina, in order to improve the competitiveness of higher education. The conclusion also points to directions for further research in this area, as well as to the limitations of the present study.

2 Literature Review

Theoretical discussions on investment in higher education are often limited to one-off discussions on the details of a particular model of investment, without considering the role and impact of higher education on overall social and economic development and competitiveness. Further, these discussions rarely perceive the performance of individual models from the perspective of a better-quality investment in higher education, nor do they consider the impact of investment on the competitiveness of higher education. Bearing this in mind, the theoretical consideration of investment in higher education begins with a review of the new trends and paradigms of higher education relating to the role, function and impact of higher education on overall social and economic development, as well as on competitiveness.

In the scientific literature, there is a number of different concepts and attitudes about the role and impact of higher education on social development, which points out the fact that there are a large number of parties interested in the optimal functioning of higher education. The views on the role and impact of higher education related to functionalism focus on outcomes in higher education. The representatives of this concept [1], [19] start from the idea of knowledge as the main element of higher education: "knowledge is material, and research and teaching are the main technologies" [1, p. 12].

Representatives of the functional approach to higher education believe that the choice of the most appropriate model of investment in higher education is greatly influenced by the fact that changes and integrations are the most important challenges of higher education. Investment models can be chosen so that the speed of change and the integration of higher education are improved. However, the extent to which a particular investment model can be applied in a given system of higher education is primarily determined by the required measure of change of the education system.

In academic debates regarding the most appropriate investment model in higher education, there exist both instrumentalist and utilitarian attitudes towards higher education, albeit reductionist regarding the understanding of social reality [11], [20], [6]. Arguments in favour of the application of one of the investment models are based on the expected or actual impact of higher education on social, economic and personal development, particularly in terms of the share of public and private investment in higher education. Authors who advocate for public investment claim that, if it is assumed that employees with a university degree earn more than those who have not attained university degrees, in the case of progressive tax rates on income, employees with a university education will be paying higher taxes. If the future revenue on the grounds of such tax is more than current expenditure, increased public investment in higher education is automatically justified. On the other hand, investment models in higher education where the primary source of income comes from public funds are not considered suitable investment models in higher education by some authors [20], [6]. According to them, such models do not contribute to the prosperity, development and increase of the competitiveness of higher education, and therefore investment models which include both public and private sources of funding are preferred. Del Rey and Racionero [6] believe that a successful and competitive investment in higher education should involve investments from two sources: public, through tax subsidies where higher education costs are paid for from general taxation, and private, where every student pays for their education through loans. They advocate joint investment in higher education, from public and private sources, without stating the optimal proportion of such investment [15].

In the area of economic development, higher education increases productivity and competitiveness, particularly through the growth of the human capital and the creation of a better educated, more qualified and more skilled workforce. In addition, in the knowledge economy, knowledge production, together with the effective and efficient transfer of knowledge to industry (in the broad sense), is one of the key factors for economic growth [13], [4], [17]. The connection between education and economic growth is suggested by other studies [16], with the conclusion that an additional year of schooling leads to an increase in GDP per capita of 4%-7%.

Considering the problem of the competitiveness of higher education, some authors [7], [23] state that internationally oriented universities contribute to the competitiveness of countries and thus determine whether countries benefit from globalization, and in what way. A major obstacle to the greater competitiveness of higher education is outdated academic structure and organization of universities, rendering them unable to guarantee academic prestige. Academic competitiveness can be improved by greater investment in higher education, opening national programs of university funding to foreign participants, as well as intensifying transnational programs of higher education financing.

The amount of tuition paid by students of higher education institutions affects the competitiveness of higher education. In many countries, tuition, i.e. the cost of higher education covered by students, is regulated by the state through a model of investment in higher education. By keeping tuition costs at specified levels, the state tries to make higher education more accessible, and this is the main argument for the implementation of the policy of regulating the price of tuition. Empirically, it turns out, however, that the price flexibility of demand of higher education is elastic [5], meaning that if there is no response to changes in tuition by students, the increase in demand that occurs as a result of regulated tuition fees is limited.

Development trends and tendencies in the field of higher education suggest that the development of the competitiveness of higher education should be directed towards more differentiated higher education systems, especially when there is a broad consensus that it is necessary to establish the country's elite universities. The fact is that the dual objective of mass access and excellence requires a dynamic and competitive sector of higher education [2]. Globalization expands the market of higher education beyond state borders, which increases the competition and demand for higher education. In order to help differentiate on the basis of the quality of higher education, it is believed [12] that higher education institutions should be allowed to choose their own tuition rates.

The analysis of the consequences when institutions are allowed to choose their own fees is particularly accentuated, because the fees in such case would represent actual costs and market situation of higher education. This would certainly encourage competition in the market for higher education. Universities would make an effort to stand out by providing a range of specific programs of study, i.e. a specific combination of price and quality [3]. The difference between supply and demand would be levelled as institutions would become more aware of the real needs of the students and their social demands. This would also encourage competition to attract students (by giving a discount on tuition fees for example), and institutions would try to win over students whose profiles best match the specific study programs in their offering.

It is necessary, however, to note two facts [5]: firstly, there is the possibility of expressing the view that the functioning of any institution of higher education cannot be the same as other companies that operate in the normal open market. While commercial companies mainly operate on the principle of maximizing the profits of their investors, institutions of higher education are mostly looking for the highest quality and for academic reputation. In this regard, it is difficult to expect that the pricing policy of education and tuition, i.e. fees dictated by the higher education institutions, would strictly follow the principle of profit maximization. Secondly, the planned consequences of the deregulation of the prices of tuition fees in higher education would come into play only when the market of higher education is completely free.

In contrast to the abovementioned studies, which are listed as comparative or similar studies in the area of research, this study deals with the econometric analysis of the competitiveness of higher education in Serbia and in AP Vojvodina, in relation to the competitiveness of 26 countries of the European Union, Croatia and the FYR of Macedonia in the period from 2006 to 2010.

3 Research Methodology

The competitiveness index of higher education in the 26 studied countries of the European Union, Serbia and its AP Vojvodina, as a European region, Croatia and FYR Macedonia, is calculated using the modified method modelled on the research methodology of calculating the index of competitiveness of European countries / regions created by Robert Huggins [21]. The aforementioned methodology was originally applied to rank the regions in the UK, as well as the countries / regions in the European Union. The period investigated in this paper is from 2006 to 2010, and the methodology applied involves the use of methods of analysis and synthesis and the use of statistical and mathematical methods, particularly data envelopment analysis (DEA method), which is the application of linear programming method based on the ranking of countries / regions according to individual indicators analysed.

Determining the competitiveness index of the higher education of these countries and AP Vojvodina involves the analysis of the three indicators which are, according to the applied research methodology, in function of the change of competitiveness of higher education:

- The number of students per 1000 population (P_1);
- 100 - the number of students per 100 employees (P_2), and
- Budgetary resources per student in EUR (P_3).

The study used a modified indicator: 100 – the number of students per 100 employees, in order to avoid the effect of favouring the country / region that has a high unemployment rate.

The DEA method determines the competitiveness index of the higher education of the countries / regions analysed. There are n elements (countries) E_i, i = 1,2, ..., n with k parameters each P_{ij}, j = 1,2, ..., k. The condition of using the calculation methods and the correct interpretation of the results is that the data are consistent, i.e., that the higher value of each indicator has the same positive (or negative) effect on the common indicator; otherwise, it is necessary to calculate the individual indices through the reciprocal value, opposite number or subtraction of a constant.

The rank of each element is identified by each indicator R_{ij}, so that the element with the highest value of the indicator has the rank of 1 and there is an adequate number of points assigned to each rank, so that the element with the highest value of indicators has the highest number of points:

$$Y_{ij}=n+1-R_{ij}$$

(1)

where:

Y_{ij} : The number of points of element rank

R_{ij}: Element rank

n : Number of elements

Proxy variables v_j for each indicator are introduced and a set of n constraints is formed:

$$\sum_{j=1}^{k} R_{ij}v_j \leq 1, \quad i=1,2,...,n \tag{2}$$

where:

k : Number of indicators

v_j : Dummy variables which correspond to certain indicators.

provided that:

$$0 \leq v_j \leq 1, \ \forall j \tag{3}$$

Linear programming, i.e. the DEA method, is used for determining the optimum value of the dummy variable with function criteria:

$$\max \sum_{j=1}^{k} R_{ij}\cdot v_j = z_{i(i)} \quad i=1,2,...,n \tag{4}$$

This gives an n set of optimal values of dummy variables $\left(v_{i1}, v_{i2}, ..., v_{ik}\right)$, for which the derived indicators are calculated for each element of the set:

$$z_{h(i)}=\sum_{j=1}^{k} R_{hj}\cdot v_{hj}, \quad i,h=1,2,...,n \tag{5}$$

where:

$z_{h(i)}$: Derived indicator

v_{hj}: Dummy variables that correspond to certain indicators.

The average value of the derived indicators for each element is the geometric mean:

$$D_i = \left(\prod_{h=1}^{n} z_{h(i)} \right)^{\frac{1}{n}} \tag{6}$$

With the explanation:

D_i : Geometric mean

$z_{h(i)}$: Derived indicator

Standardization of the derived indicators is performed:

average value: $\bar{D} = \frac{1}{n} \cdot \sum_{i}^{n} D_i$ (7)

deviation: $\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(D_i - \bar{D} \right)^2}$ (8)

standardized value: $s_i = \frac{D_i - \bar{D}}{\sigma}$ (9)

Calculating the average deviation of the initial values of indicators:

The average value of the indicator: $\bar{P}_j = \frac{1}{n} \sum_{i}^{n} P_{ij}$ (10)

$\bar{P}_j$: Average values of original indicators

P_{ij} : Original value of indicators

The deviation of the initial values of parameters is adjusted so that the average value is 100:

$$\sigma_j = \left[\frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{P_{ij}}{\bar{P}_j} \cdot 100 - 100 \right)^2 \right]^{\frac{1}{2}} \tag{11}$$

The average deviation of the original value of the indicator is calculated as the geometric mean to the formula:

deviation of the average value: $\sigma_p = \left(\prod_{j=1}^{k} \sigma_j \right)^{\frac{1}{k}}$ (12)

Standardized derived indicators are corrected with an average deviation of the initial values of indicators:

$$K_i = S_i \cdot \sigma_p + 100 \tag{13}$$

where:

K_i : Final value of DEA index points.

The obtained values of the DEA index points K_i indicate the relative position of a given element in the set and are in line with the size of the original variables. The meaning of the points obtained is in accordance with the meaning of initial variables; i.e., if the initial variables indicated a better position of the elements of a set, it was shown by the obtained points.

The study, using the described methodology first, and based on initial data for the given indicators, calculated the DEA indices of the competitiveness of the higher education of 26 countries of the European Union, Serbia, AP Vojvodina (as a European region), Croatia and the FYR of Macedonia for the period from 2006 to 2010. The indices for Serbia and AP Vojvodina were calculated by the current model of investment in higher education [14]. The current model of investment is based on a formula with exclusive entrance criteria (the number of students enrolled for the first time, the number of teaching and non-teaching staff) in the allocation of budgetary resources for public institutions of higher education.

Thereafter, the authors applied a new investment model in higher education [15] to the baseline data for these indicators of competitiveness, which refer to Serbia and AP Vojvodina, ex post, in the same conditions. A new investment model in higher education is based on a formula that contains a combination of input - output criteria (the number of first time enrolments and the number of graduates, the proportional cost per student) in the allocation of investment, budgetary resources per student in study programs in higher education. By calculating investment costs per student in a study program and increasing the number of state-funded students, a new investment model, in the long term, encourages the employment of graduates. The faster employment of graduates of higher education through planning is a qualitative feature of the new model of investment in higher education relating to the results of its application. The aim and purpose of the new model of investment is the analysis of its influence on the changes in the competitiveness ranking of higher education of Serbia and AP Vojvodina, as a European region.

4 The Research Results

4.1 The DEA Index of Competitiveness of Selected Countries in Europe, Serbia and AP Vojvodina for the Period from 2006 to 2010

Using the described methodology, based on the initial indicators of competitiveness, the DEA index of competitiveness of higher education and their ranking was calculated for the 26 countries of the European Union, Serbia, AP Vojvodina (as a European region), Croatia and FYR Macedonia for the period from 2006 to 2010, and the data are shown in Table 1.

Table 1

DEA competitiveness index of higher education and ranking of selected European countries, Serbia and AP Vojvodina, in the period from 2006 to 2010

No		2006		2007		2008		2009		2010	
	Country /Region	DEA index	Rank	DEA index	Rank	DEA index	Rank	DEA index	Rank	DEA index	Rank
1	Serbia	57.58	30	56.60	30	59.43	30	56.19	30	55.52	30
2	AP Vojvodina	59.92	29	61.74	29	64.51	29	64.17	29	59.95	29
3	Croatia	86.07	26	87.87	26	90.74	24	92.06	23	88.27	26
4	FYR Macedonia	70.61	28	71.50	28	69.38	28	76.01	28	77.45	28
5	Belgium	98.98	19	97.64	20	99.24	20	99.51	19	103.04	12
6	Bulgaria	94.67	21	99.51	19	102.67	16	103.46	14	98.63	20
7	Czech Rep.	108.16	12	109.10	13	109.19	10	110.62	9	109.74	11
8	Denmark	121.52	1	116.69	1	116.84	2	117.36	2	117.05	3
9	Germany	104.64	15	107.10	14	111.32	6	114.38	5	113.04	6
10	Estonia	112.49	8	111.94	7	108.04	13	101.23	17	99.92	18
11	Ireland	114.90	4	112.29	6	109.37	9	103.53	13	102.03	14
12	Greece	89.54	25	89.08	25	86.81	27	87.73	27	88.20	27
13	Spain	107.99	13	106.10	15	104.16	15	100.11	18	99.82	19
14	France	98.25	20	97.56	21	100.18	19	102.30	16	100.97	16
15	Italy	93.98	22	93.40	23	94.97	23	96.44	22	95.91	23
16	Cyprus	105.07	14	109.52	11	111.46	5	112.90	7	113.51	5
17	Latvia	111.53	9	112.82	5	108.67	12	96.98	20	97.17	21
18	Lithuania	102.38	16	103.19	16	96.81	21	89.78	26	90.41	25
19	Hungary	91.36	23	89.17	24	88.21	25	89.91	25	91.87	24
20	Malta	89.89	24	95.67	22	101.67	18	104.14	12	102.77	13

21	Netherlands	116.64	3	115.57	2	117.20	1	120.28	1	118.37	1
22	Austria	110.17	11	110.52	9	112.78	4	115.29	3	115.68	4
23	Poland	81.80	27	86.71	27	87.36	26	90.87	24	96.81	22
24	Portugal	111.23	10	110.54	8	110.88	7	112.50	8	110.52	9
25	Romania	100.04	18	99.90	18	95.77	22	96.84	21	100.02	17
26	Slovenia	114.38	5	113.05	4	109.07	11	109.44	10	111.05	8
27	Slovakia	100.59	17	101.29	17	102.24	17	102.48	15	101.91	15
28	Sweden	118.83	2	114.55	3	114.05	3	114.40	4	117.52	2
29	United Kingdom	114.20	6	110.20	10	110.56	8	113.27	6	112.55	7
30	Finland	112.59	7	109.19	12	106.43	14	105.83	11	110.31	10

Source: Eurostat (2010), Statistical Office of the Republic of Serbia (2011), and author's own calculations

Based on the data in Table 1, it can be concluded that the greatest competitiveness of higher education among the 30 surveyed countries / regions in Europe is in countries where the output criteria and factors involving increased amounts of budget investment are the basis of models of investment in higher education. Denmark is in the first place in 2006 and 2007, and the Netherlands is in the first place in the next three years analysed. Listed after them is another Scandinavian country, Sweden, which in the given five-year period is in the second or the third place, except in 2009 when the country was listed in the fourth place. In Sweden, in recent years, stricter connections were enforced between academic progress and budget grants, which is favourable to the further development of higher education and its competitiveness.

Among the other highly industrialized countries, there is significant competitiveness of higher education in the five-year period analysed, in the cases of Great Britain, Austria and Germany. In these countries, the increased demand for higher education scholarships was matched by higher tuition prices. A special feature of these industrialized countries is that in the last three years analysed, i.e. in 2008, 2009 and 2010, they significantly improved the competitiveness of higher education, analysed using the competitiveness index. On the other hand, two highly developed industrial countries of the European Union, Italy and France, have low competitiveness of higher education, primarily due to lower budget allocations per student compared to other developed countries of the European Union. This especially applies to Italy, which was located at the 22nd or 23rd place in the five-year period, out of the 30 European countries / regions analysed.

It is important to mention the index of higher education competitiveness of Ireland and Cyprus, which are seeking to step up their social and economic development based primarily on the principle of the "knowledge society" [18]. While in Cyprus, competitiveness was on the rise in the period from 2006 to 2010, in Ireland, affected by the economic crisis, it was quite the opposite: there the competitiveness of higher education decreased in the five-year period analysed.

The lowest competitiveness of higher education among the analysed countries of the European Union was found in Greece, in the 25th position in 2007 and 2006 and in the 27th position in the other three years of the five-year period. The index points of the most competitive countries and of the countries with the lowest competitiveness indicate that there are clearly noticeable disparities in competitiveness of higher education among the analyzed countries / regions, because the difference is more than 60 percentage points.

In the countries of former Yugoslavia, Slovenia was highly ranked in all years of the five-year period. This especially applies to 2006 when it ranked fifth, and even more so in 2007 when it ranked fourth. Croatia ranks low in 2006, 2007 and 2010, at the 26th place, in 2008 at the 23rd place and in 2009 at the 24th place in the five-year period analysed. The lowest competitiveness of higher education among the countries of former Yugoslavia, according to the methodology applied, is found in FYR Macedonia. While in Slovenia the applied investment model in higher education is characterized by a high degree of financial autonomy of institutions, as well as a combination of input and output criteria in the model formula, in the current investment model in Serbia, as well as the investment model of Croatia and FYR Macedonia, there is a low level of financial autonomy of institutions and mostly input criteria in the allocation of budgetary funds. These facts, related to the performance of the investment model, affect the competitiveness of the higher education of these countries.

Because of the least favourable initial values of the indicators of the competitiveness of higher education, the competitiveness of the higher education of Serbia and AP Vojvodina is the lowest of all the European countries analysed. According to the competitiveness of higher education, analyzed using the existing model of investment in which the criteria are based on inputs, Serbia is at the 30th place in all years of the five-year period, and AP Vojvodina is at the penultimate 29th place. This data indicates the need to implement a new investment model for the higher education system in Serbia to improve investment and to increase the number of state-funded students and graduates employed, and thus to contribute to an increase of higher education competitiveness.

4.2 DEA Competitiveness Index, Serbia and AP Vojvodina, for the Period from 2006 to 2010 after the Implementation of the New Investment Model

Based on the applied methodology, the DEA indices were calculated and the rankings of the competitiveness of higher education in Serbia and AP Vojvodina were determined and compared to the 26 countries of the European Union, Croatia and FYR Macedonia, after the implementation of the new investment model [15]. These are shown in Table 2.

Table 2

DEA competitiveness index of higher education and ranking of selected European countries, Serbia and AP Vojvodina, in the period from 2006 to 2010 after the implementation of the new investment model

No		2006		2007		2008		2009		2010	
	Country /Region	DEA index	Rank	DEA index	Rank	DEA index	Rank	DEA index	Rank	DEA index	Rank
1	Serbia	61.18	30	57.75	30	62.70	30	59.41	30	58.72	30
2	AP Vojvodina	62.94	29	62.71	29	67.45	28	66.94	29	62.68	29
3	Croatia	84.93	26	87.50	26	89.94	24	91.33	23	87.24	26
4	FYR Macedonia	68.63	28	70.79	28	67.41	29	74.38	28	75.74	28
5	Belgium	98.53	19	97.46	20	98.86	20	99.13	19	102.85	12
6	Bulgaria	94.01	21	99.38	19	102.50	16	103.33	14	98.20	20
7	Czech Rep.	108.27	12	109.18	13	109.36	10	110.86	9	109.98	11
8	Denmark	122.33	1	116.91	1	117.39	2	117.96	2	117.71	3
9	Germany	104.58	15	107.16	14	111.70	6	114.92	4	113.57	6
10	Estonia	112.74	8	112.03	7	108.03	13	100.85	17	99.44	18
11	Ireland	115.33	4	112.40	6	109.52	9	103.37	13	101.77	14
12	Greece	88.43	25	88.66	25	85.61	27	86.56	27	86.93	27
13	Spain	108.04	13	106.10	15	104.04	15	99.76	18	99.43	19
14	France	97.78	20	97.40	21	99.89	19	102.13	16	100.70	16
15	Italy	93.27	22	93.14	23	94.39	23	95.94	22	95.34	23
16	Cyprus	105.04	14	109.63	11	111.80	5	113.28	7	113.98	5
17	Latvia	111.68	9	112.91	5	108.64	12	96.31	20	96.47	21
18	Lithuania	102.00	16	103.06	16	96.10	21	88.65	26	89.21	25
19	Hungary	90.44	23	88.79	24	87.19	25	88.99	25	90.99	24
20	Malta	89.02	24	95.50	22	101.56	18	104.15	12	102.70	13
21	Netherlands	117.21	3	115.78	2	117.82	1	121.08	1	119.16	1
22	Austria	110.40	11	110.64	9	113.17	4	115.83	3	116.31	4
23	Poland	80.26	27	86.23	27	86.16	26	89.85	24	96.08	22
24	Portugal	111.50	10	110.65	8	111.17	7	112.90	8	110.84	9
25	Romania	99.65	18	99.75	18	95.10	22	96.21	21	99.54	17
26	Slovenia	114.69	5	113.13	4	109.05	11	109.47	10	111.22	8
27	Slovakia	100.25	17	101.18	17	101.99	17	102.23	15	101.62	15
28	Sweden	119.46	2	114.71	3	114.42	3	114.80	5	118.18	2
29	United Kingdom	114.62	6	110.29	10	110.79	8	113.72	6	112.96	7
30	Finland	112.80	7	109.19	12	106.26	14	105.66	11	110.43	10

Source: Eurostat (2010), Statistical Office of Serbia (2011), and author's own calculations

Based on the data presented in Table 2, it can be concluded that with the implementation of a new investment model, the competitiveness of higher education of Serbia is increased, compared to the selected European countries in each year of the analysed period of five years: in 2006 by 3.60 percentage points, in 2007 by 1.15, in 2008 by 3.27, in 2009 by 3.22 and in 2010 by 3.20 percentage points.

In AP Vojvodina there are similar trends, but with lower growth of the competitiveness index of higher education. The competitiveness of higher education in AP Vojvodina, as a European region, with the implementation of a new investment model in higher education is also improved, because the initial index points using the new model of investment in 2006 are higher by 3.02, in 2007 by 0.97, in 2008 by 2.94, in 2009 by 2.77 and in 2010 by 2.73 percentage points. It is noted that, due to the low budgetary investment in higher education in AP Vojvodina and the fewer students per 1,000 residents, by using the new investment model and methodology described, there was a lower pace of the growth of the competitiveness index of higher education in relation to the growth of the same index in Serbia as a whole.

According to the new model of investment, there was no change in the rank of Serbia in relation to the selected European countries and AP Vojvodina, as a European region, since in all years of the five-year period analysed, the competitiveness ranking of Serbia remained at the 30th place. It is similar with AP Vojvodina, which, using the new investment model in higher education and compared to selected European countries improved its competitiveness ranking in 2008 only.

Based on these data and facts, it is shown that the application of the new investment model contributed to increasing the competitiveness of higher education in Serbia and in AP Vojvodina, and its application is necessary and expedient.

Conclusion

The results of the analysis presented in this paper suggest an unsatisfactory competitiveness of higher education of Serbia and AP Vojvodina, as a European region. The competitiveness is not adequate to the need and dynamics of development of higher education in transition countries like Serbia. Stronger social and economic development is conditioned by creating a knowledge based society, and by learning and innovation, which includes the need for further development of higher education in Serbia and an increase of its competitiveness. Therefore, the higher education system should direct its educational and research capacity at increasing development needs of society and the economy.

Facts and figures obtained in this study suggest the need to intensify efforts and resources in the field of higher education with the aim of increasing the competitiveness of higher education in Serbia and AP Vojvodina, in relation to the countries of the European Union. On the basis of these results, the educational authorities in the higher education of Serbia are suggested to implement a new investment model. The application of the new investment model would increase the volume of investment in higher education, the number of state-funded students, the employability of graduates, and thus the competitiveness of higher education.

The results indicate that a unified approach to strategy development and higher education competitiveness in the European Union is necessary. This is because there are very large differences in the competitiveness of higher education among the European countries analysed. The creation of a knowledge based society implies a higher share of highly educated population. Thus, the need to increase the quality and quantity as well as the international competitiveness of higher education is one of the basic issues for which authorities in the higher education sector in European countries should find appropriate solutions in the future.

Restrictions regarding the research are related to the fact that this research did not analyse the higher education competitiveness of the EU as a whole. There are also limitations regarding the indicators of competitiveness of higher education used to calculate the DEA competitiveness index. Some factors which are not only in the spheres of policy makers and development strategy of higher education affect the change of these indicators, such as demographic changes, the socio-economic status of the population, the interest in studying in the function of labour market trends, and the like. On the other hand, there is the issue of unemployment or employment reduction due to the effects of the economic and debt crisis and its impact on the analysed indicators.

Bearing this in mind, the directions for further research will be related to the DEA analysis of the competitiveness index of higher education by researching other relevant indicators of competitiveness of higher education, such as the number of graduates in the total number of students and higher education participation in total employment.

References

[1] B. Clark: The Higher Education System: Academic Organization in Cross National Perspective, Berkeley, University California Press, 1983

[2] B. Jacobs, F. Ploeg: Guide to Reform of higher Education: A European Perspective, Discussion paper series No. 5327, Industrial organization and public policy, Centre for Policy Research, London, 2005

[3] C. M. Hoxby: How the Changing Market Structure of U.S. Higher Education Explains College Tuition, NBER, Working Paper 6323, 1997

[4] C. Lentner: The Competitiveness of Hungarian University-based Knowledge Centres in European Economic and Higher Education Area, Transformations in Business & Economics, Vol. 6, No. 2 (12), 2007, pp. 87-99

[5] E. Canton, et al: Higher Education Reform: Getting the Incentives Right. CPB/CHEPS University of Twente, 2001

[6] E. Del Rey, M. Racionero: Financing Schemes for Higher Education, European Journal of Political Economy, Vol. 26, No. 1, 2010, pp. 104-113

[7] E. Egron – Polak: Which Way Ahead for IAU? The Bimonthly Newsletter of the International Association of Universities, Vol. 7, No. 5, 2001, pp. 1-3

[8] Eurostat: Employment statistic, 2010
http://epp.eurostat.ec.europa.eu/portal/page/portal/employment_unemployment_lfs/data/database

[9] Eurostat:Population statistics, 2010
http://epp.eurostat.ec.europa.eu/portal/page/portal/population/data/database

[10] Eurostat:Tertiary education statistic, 2010
http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Tertiary_education_statistics

[11] J. Eicher: The Costs and financing of Higher Education in Europe, European Journal of Education, Vol. 33, No. 1, 1998, pp. 31-39

[12] J. Eicher, T. Chevailler: Rethinking the Financing of Post-Compulsory Education. Higher Education in Europe, Vol. 27, No. 1-2, 2002, p. 69

[13] J. G., Mora, L. Vila: The Economics of Higher Education. In: The Dialogue between Higher Education Research and Practice, ed. Roddy Begg, 121-134, Kluwer Academic Publishers, 2003

[14] J. Kabók: Investment Models for Higher Education, Annals of the Faculty of Economics in Subotica, Vol. 46, No. 24, 2010, pp. 155-168

[15] J. Kabók, V. Djaković, G. Andjelić: Investment Model Aimed at Raising Competitiveness of Higher Education, Proceedings XV International Scientific Conference on Industrial Systems, Novi Sad, Serbia, September 14-16 2011, pp. 446-450

[16] J. Lowter: The Quality of Croatian Formal Education System. In: Competitiveness of Croatian workforce, Institute of Public Finance, Zagreb, 2004, pp. 13-25

[17] K. Dobrai, F. Farkas, Zs. Karoliny, J. Poór: Knowledge Transfer in Multinational Copmanies – Evidence from Hungary, Acta Polytechnica Hungarica, Vol. 9, No. 3, 2012, pp. 149-161

[18] Lisbon European Council 23 and 24 March 2000: Presidency Conclusions, 2000
http://www.europarl.europa.eu/summits/lis1_en.htm#1

[19] L. Weber, S. Bergan: The Public Responsibility for Higher Education and Research, Council of Europe Higher Education Series No. 2, Council of Europe, Strasbourg, 2005

[20] N. Barr: Higher Education Funding, Oxford Review of Economic Policy, Oxford University Press, Vol. 20, No. 2, 2004, pp. 264-283

[21] R. Huggins: Designing a European Competitiveness Index: Measuring the Performance and Capacity of Europe's Regions and Nations', European Regional Economic Forum, Nova Gorica, Slovenia, October 27-28, 2005

[22] Statistical Office of the Republic of Serbia, Statistical Almanac of Serbia, Belgrade, 2010

[23] T. Berchem: The University as an Agora – Based on Cultural and Academic Values. Higher Education in Europe, Vol. 31, No. 4, 2006, pp. 395-396

[24] The Bologna Declaration of 19 June 1999: Joint declaration of the European Ministers of Educatin, 1999

http://www.bologna-bergen2005.no/Docs/00-Main_doc/990719BOLOGNA_DECLARATION.PDF

[25] UNESCO Access, Values, Quality and Competitiveness, UNESCO Forum on Higher Education in the Europe Region, Bucharest, Romania, May 21-24, 2009

Multi-Paradigm Metric and its Applicability on JAVA Projects

Sanjay Misra

Department of Computer Engineering, Atilim University
Kizilcaşar Mh., 06830 Incek, Ankara/Ankara Province, Ankara, Turkey
smisra@atilim.edu.tr

Ferid Cafer

BOTT Information Systems, Silicon Block No:20
Middle East Technical University Teknopolis, 06531, Ankara, Turkey
ferid.cafer@bott.com.tr

Ibahim Akman

Department of Computer Engineering, Atilim University
Kizilcaşar Mh., 06830 Incek, Ankara/Ankara Province, Ankara, Turkey
akman@atilim.edu.tr

Luis Fernandez-Sanz

Universidad de Alcalá, Depto. de Ciencias de la Computación
Plaza de San Diego, s/n, 28801 Alcalá de Henares, Madrid, Spain
luis.fernandezs@uah.es

Abstract: JAVA is one of the favorite languages amongst software developers. However, the numbers of specific software metrics to evaluate the JAVA code are limited. In this paper, we evaluate the applicability of a recently developed multi paradigm metric to JAVA projects. The experimentations show that the Multi paradigm metric is an effective measure for estimating the complexity of the JAVA code/projects, and therefore it can be used for controlling the quality of the projects. We have also evaluated the multi-paradigm metric against the principles of measurement theory.

Keywords: Multi-paradigm metric; Software complexity; JAVA; software development

1 Introduction

One of the important issues in the software development process is to maintain the quality of the software. Complex codes are not desirable because they are hard to maintain and reduce the quality of the software [1] [2]. The complex codes also decrease the understandability and increase the burden on reviewers, testers and maintainers. In this point of view, if the complexity is not controlled from the beginning of software development, it may cause higher maintainability and reduce the quality of the product. As a result, complex code increases the cost of software/software product. To overcome this issue, the complexity of the code should be controlled. Software metrics are the tools to control the complexity. Researchers are making continuous efforts to produce metrics to control the complexity of the code. Further, software metrics tend to compare various parameters such as cost, effort, time, maintenance, understanding and reliability. Metrics are indispensable from several aspects, such as measuring the understandability of a code, the testability of the software, the maintainability and the development processes [3].

Over last two decades, object oriented programming languages have gained considerable acceptance from the software development community. Among several object-oriented languages, JAVA has become a favorite language for developing software products. The popularity of JAVA has arisen as a consequence of its unique features. On the other hand, to evaluate the quality of the software code written in JAVA, few metrics [4] are available in the literature. We mention the effort of researchers [4-10] who tried to control the quality of JAVA by considering different aspects and features of JAVA programming. Dufour [4] proposed dynamic metrics for JAVA. Cahoon et al. [5] worked on data flow analysis for software perfecting linked data structures in JAVA controllers. Sudaresan et al. [6] researched practical virtual method call resolution for JAVA. Vijaykrishnan et al. [7] have produced tuning branch predictors to support virtual method invocation in JAVA. Qian et al. [8] proposed a comprehensive approach to array bounds check elimination for JAVA. Erik Ruf [9] proposed a methodology for the effective synchronization removal for JAVA. Shuf et al. [10] proposed a structured view and opportunities for optimizations by characterizing the memory behaviour of JAVA workloads. Mäkelä et al. [11] proposed a new client based metric, Lack of Coherence in clients (LCIC), and developed a tool for measuring the metric for JAVA projects. The authors tried to improve the quality of code through the LCIC metric, which measures how a class is used by other classes in a context. In their comparison analysis, the authors also suggested which type of refactoring is required. In an another empirical study of JAVA inheritance evaluation, Nasseri et al. [12] found that larger and highly coupled classes were less cohesive and more frequently moved than smaller and less coupled classes. Kaczmarek and Kucharski [13] demonstrated how to estimate size and efforts for JAVA based applications. They presented three models of size estimations, which were based on class and method size. The authors concluded that for big projects, for example projects of nearly one million lines of code, the average class and method size will be independent from the application size. Giuseppe [14] proposed a semantic similarity metric which combines the features and intrinsic information content. Romano [15] proposed using source code metrics to predict change-prone JAVA interfaces. None of the above works evaluate the quality of the JAVA code, which is responsible for the understandability and therefore the maintainability of the JAVA product. It is worth mentioning that maintainability is identified as one of the most important software quality [16] attributes.

In this paper, we apply a recently proposed metric [17] that was developed for multi-paradigm languages on JAVA projects. A multi-paradigm language is a language which includes features of two or more than two programming paradigms. The metric developed in [17] considered procedural, object oriented paradigm, and combined the function point metric [18] to estimate complexity due to the functionality of the code/project and an object oriented metric [19] to estimate the complexity of object oriented features. The proposed multi-paradigm metric [17] was applied in a project written in C++. Since JAVA is also a multi paradigm language, which includes features of procedural and object-oriented language, we apply the same metric to evaluate JAVA projects. In fact, the agenda of the present paper is twofold. Firstly, we want to check the applicability of the multi-paradigm metric in JAVA projects; and secondly, we want to perform more experimentation for the empirical validation of the multi paradigm metric, given that the real applicability of a metric cannot be proved without a series of empirical observations [20]. Even more, we will evaluate the theoretical soundness of the multi paradigm metric by applying the principles of measurement theory to the multi paradigm metric.

Before moving ahead, we would like to summarize our previous works in this area. We have developed metrics for procedural languages [21], object oriented languages [19], [22], [23], and multi paradigm languages [17]. One of the coauthors of this paper is also involved in developing metrics for various purposes, e.g. web-services, [24], [25], SOA and XML schema languages[26], [27], Business Process Modeling[28], etc. Another coauthor has proposed a scheme for the verification and validation of JAVA code by combining code style check and some code metrics to prioritize test cases [29].

The structure of the paper is as follows. The definition of the multi paradigm metric is given in Section 2. In Section 3, we first evaluate the metric to check its soundness from the principles of measurement theory, and then we apply the metric on JAVA projects in Section 4. The conclusion of the work is in the last section.

2 Multi-Paradigm Complexity Measure Measurement

In order to compute the complexity of software system, the authors [17] have suggested how to compute the quality of the code by considering that the overall complexity of the software system depends on the functionality as well as on different factors of the object oriented and procedural parts of the system.

Accordingly, the computation of the quality of code for multi-paradigm programs is presented as,

Code Quality (CQ): The CQ is defined by the number of function points to the complexity values due to all the factors in the multi-paradigm program code.

CQ = (FP / MCM)* 10,000 (1)

where, FP [30] is the Function Point calculations for the code, and MCM represents the multi-paradigm complexity measurement and computes the complexity of the code as given in equation (2). MCM followed the similar approach of a metric developed for python [13].

$$MCM = CIclass + CDclass + Cprocedural \qquad (2)$$

where **CIclass** = Complexity of Inherited Classes,

CDclass = Complexity of Distinct Class,

and **Cprocedural =** Procedural Complexity.

All these factors are defined as follows:

The complexity of an independent class is calculated first because it plays a role either in the inheritance hierarchy or as a distinct class. In other words, for calculating CIclass or CDclass, first it is necessary to calculate Cclass, the complexity of an independent class. The complexity (Cclass) of an independent class can be computed as:

$$Cclass = W(attributes) + W(variables) + W(structures) + W(objects) - W(cohesion) \quad (3.1),$$

where **Cclass** represents the Complexity of a single class.

In the above formula, the weight due to cohesion is subtracted because it reduces the complexity and is desirable from the point of view of software developers [1].

The weight of attributes or variables is computed as:

$$W(variables\ or\ attributes) = 4 * AND + MND \qquad (3.1.1)$$

where AND represents the Number of Arbitrarily Named Distinct Variables/Attributes, and

MND represents the Number of Meaningfully Named Distinct Variables/Attributes.

Weight of structure: W (structures) is defined as the weight of structure of the methods inside the class:

$$W(structures) = W(BCS) \tag{3.1.2}$$

where BCS are basic control structures.

Weight of objects Weight (objects) is computed as:

$$W(objects) = 2 \tag{3.1.3}$$

The weights of objects are assigned as 2, because it is similar as to how an object constructor is automatically called while creating it and it is a coupling. In other words, calling a function or creating an object represents the same complexity. The coupling can also occur due to method calls, which are already considered while computing the weight of structure in MCM.

Weight of cohesion is defined as:

$$W(cohesion) = MA / AM \tag{3.1.4}$$

where MA represents the Number of methods where attributes are used, and

AM represents the Number of attributes used inside methods.

While counting the number of attributes, there is no any importance of AND or MND.

CIclass can be defined as:

There are two cases for calculating the complexity of the Inheritance classes depending on the architecture:

- If the classes are in the same level then their weights are added.
- If they are children of a class, then their weights are multiplied due to inheritance property.

If there are *m* levels of depth in the object-oriented code and level j has n classes, then the Cognitive Code Complexity (CCC) [23] of the system is given as

$$CIclass = \prod_{j=1}^{m}\left[\sum_{k=1}^{n} CC_{jk}\right] \tag{3.2}$$

CDclass can be defined as:

$$CDclass = Cclass(x) + Cclass(y) + \ldots \tag{3.3}$$

Note: All classes that are neither inherited nor derived from another are parts of Cdclass even if they have caused coupling together with other classes.

Cprocedural can be defined as:

$$Cprocedural = W(variables) + W(structures) + W(objects) - W(cohesion) \tag{3.4}$$

Weight of variable *W*(variable) is defined as:

$$W(varialbes) = 4 * AND + MND \quad (3.4.1)$$

The variables are defined globally.

Weight of structure *W*(structures) is defined as the weights of all the:

$$W(structures) = W(BCS) + object.method \quad (3.4.2)$$

where BCS are basic control structures, and those structures are used globally. 'object.method' is a reachable method of a class using an object. 'object.method' is counted as 2, because it is calling a function written by the programmer. If the program consists of only procedural code, then the weight of the 'object.method' will be 0.

Weight of objects W(objects) is defined as:

$$W(objects) = 2 \quad (3.4.2)$$

Creating an object is counted as 2, as is described above (3.1.3). Here it refers to the objects created globally or inside any function which is not a part of any class. If the program consists of only procedural code, then the weight of the 'objects' will be 0.

$$W(cohesion) = NF / NV \quad (3.4.3)$$

where NF is the number of functions and NV means number of variables. Coupling is added inside *W* (structures) as mentioned in the beginning of the description of the metric.

3 Theoretical Validation

For the theoretical validation of the proposed metric, we follow the properties proposed by Briand et al. [31]. Briand et al. proposed five properties for evaluating a complexity metric. These properties provide a useful guideline in the construction and validation of complexity measures and have been used by several researchers [22], [32], [33]. In the following sections, we provide all these properties and evaluate our metric against these metrics. We want to clarify that Code quality is dependent on two different complexity metrics: Function Point and multi-paradigm complexity measurement. In our formulation, Function Point calculation is estimated at the whole project level and MCM is computed at class level. The properties proposed by Briand et al. [31] evaluate complexity measures which are applied on programs/classes/modules. From this point of view, we evaluate MCM against these properties, instead of code quality of multi paradigm programs.

Property 1: *Non-Negativity*

The complexity value given by MCM for a class can never be negative. In our formulation there is only one possibility for MCM values to be negative, when the weight of cohesion will be higher than that the sum of weights all other factors of that class. This is not possible because the weight of cohesion is computed as NF/NV, and this number cannot be greater than the sum of number of methods (if we assume the weight of each method in class is one), number of attributes and other parameters like variables, etc. Hence, MCM satisfies Property 1.

Property 2: *Null Value*

It is possible that the elements of our metric will be absent from the class; in this case our metric gives a null value. See the following Table 1.

Table 1
Metric value of elements of a class

Class	att	str	var	Obj	MA	AM	Cohesion	Comp./MCM
XX	0	0	0	0	0	0	0	0

Since the proposed measures can get a null value in a class our measures satisfy the Property.

Property complexity 3 (Symmetry): By changing the order of statements, methods, attributes, and variables, there is no effect on our metric values. In other words, MCM will not change by changing the order of its elements.

Property complexity 4 (Module Monotonicity): If we add two classes then the MCM values of the combined classes will be equal to the sum of MCM values of the individual classes. In our formulation we have also considered the effect of interference; i.e. if the classes are in a hierarchy, then first we add the complexity of classes which are the same level and then multiply with its parent's class. In this case, also Module monotonicity is preserved. We can take an example of classes in our case study. We consider three classes: Figure2P, Rectangle, and Oval. Figure2P is a parent class of Rectangle and Oval classes. We add all these three classes; the complexity of Rectangle and Oval will be added first and then multiplied by the complexity of the Figure2P. According to the property of monotonicity:

The complexity of combined classes in hierarchy will be estimated by:

Figure2P * (Rectangle + Oval) = 10 * (29 + 29)
= 580 (A)

If we sum the MCM values of all in depended classes, the MCM values of combined classes are

=Figure2P + Rectangle + Oval)
= 10 +29 + 29
=68 (B)

From equation A and B it is clear that the complexity of the combined classes is always equal to or greater than the sum of the complexity of the independent classes. As these examples confirm, our metric satisfies the module monotonicity property.

Property complexity 5 (Disjoint Module Additivity): This property states that if the two classes are combined, then the combined class's complexity will equal to the sum of complexity of the independent classes. This is the property of additivity, the most important property to achieve the scale of the metric. We will take two different examples to check this property, because classes may be arranged in two ways, first in the same level and second in different levels in class hierarchy.

1. Consider the two classes at the same level. In our case study, two classes Rectangle and Oval are at the same level. Therefore, when we combine these two classes we can easily observe that the MCM values of the combined classes, i.e. 29+29 =58, will be the same as when we combine the classes independently.
2. If we combine the classes at a different level, we will also find the same result. Suppose we combine the Figure 2P- Rectangle and Figure 2P-Oval.

   The MCM values of Figure 2P-Rectangle Class= 10*29= 290

   The MCM values of Figure 2P-Ovel Class= 10*29= 290

   The sum of the these two independent classes

   = MCM values of (Figure 2P- Rectangle + Figure 2P-Ovel)

   = 290+290

   = 580 (C)

   Now we compute the complexity of combined class Figure 2P- Rectangle-Figure 2P-Ovel, which is computed as

   =Figure2P * (Rectangle + Oval) = 10 * (29 + 29)

   = 580 (D)

   From equation C and D, it is proved that MCM satisfies the additive property.

   Hence, MCM satisfies this property too.

   After satisfying all these five properties, i.e. additivity, module monotonocity, symmetry, null values and non-negative, we can conclude that our MCM is a valid and sensible measure from the theoretical point of view. Further, if a complexity metric satisfies the fifth property, then the metric is also on ratio scale. Property 5 proves that MCM satisfies the additive property and is on ratio scale.

4 Applicability of Multi-Paradigm Complexity Metric on JAVA Projects

We have selected two projects for empirical validation of our metrics. Both projects are available online. We chose online projects due to two reasons: 1. the reader may also want to analyze the projects in the same way as the author. 2. They are completed and tested projects so one can assume them without any fault. The details of both the projects are the following:

4.1 Chatting Application

This is an application developed in JAVA for chat. The program is divided into two; client-side and server-side [34]. Inside this program inheritance between classes are not used; in fact, it has a simpler structure than other compared projects though it has a higher level of functionality. Therefore, it has the highest code quality. Its number of LOC (Lines of Code) is 1208.

Firstly, we estimate the MCM and the Function points to evaluate the code quality of this project. The components of the MCM are computed and summarized in Table 2.

Table 2

Chat Application – Classes

Class	att	str	var	obj	MA	AM	cohesion	Comp.
CLIENT_INFO	2	2	0	0	1	2	0.5	3.5
MainFrame(S)	0	39	4	20	0	0	0	63
THBind	3	20	0	6	2	3	0.6	28.4
Client_P	2	102	5	14	2	2	1	122
MSG_RDR	0	8	0	6	0	0	0	14
S_Client	0	2	0	2	0	0	0	4
MainFrame(C)	3	30	4	16	1	3	0.3	52.7
Form	3	68	0	20	2	4	0.5	90.5
Sign_UP	0	18	0	16	0	0	0	34
Frame3	0	15	0	10	0	0	0	25
CHAT_WIN	2	16	0	12	2	2	1	29
MSG_READER	0	8	0	2	0	0	0	10
CMD_L	1	34	0	2	2	1	2	35

A graph of the complexity values (MCM) for all the classes are shown in Figure 1. If we analyze this project (see Figure 1), we can find out that the maximum complexity is 122 which belongs to Client_P Class. This class is the most complex because it has the highest number of strings (22) and variables (5). In other words, this class has several control structures with variables. The average complexity of the classes of this project is 39. The least complex class is CLIENT_INFO with a complexity of 3.5. This class includes only two attributes and two strings.

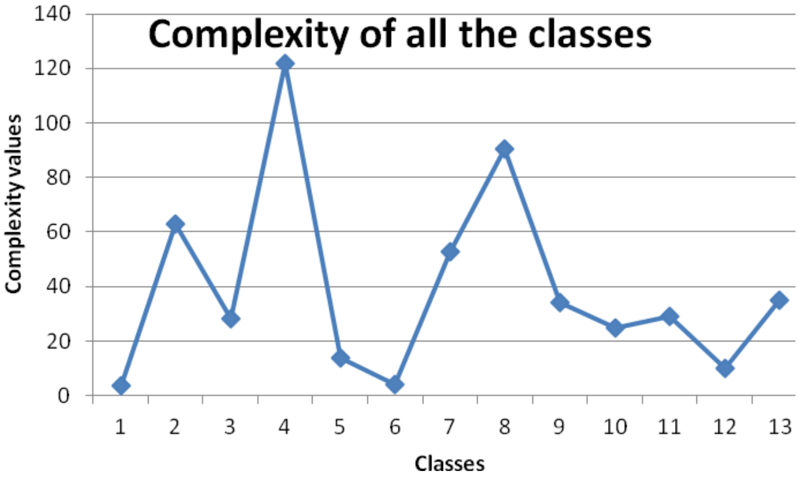


Figure 1
Complexity of Various Classes of the Chatting Application

The procedural complexity of this project is summarized in Table 3.

Table 3
Chat Application – Cprocedural

Non-Class	var+str+obj	Complexity
Cprocedural (S_CHAT)	9	9

The main program is not very complex and it consists of 9 variables, strings and the object (Table 3). Therefore, its complexity is 9 (Cprocedural).

Additionally in this project, all the classes are independent and no hierarchy amongst the classes is present. So this project

1. Does not have complexity due to inheritance. i.e CIclass = 0.
2. All the classes are treated as distinct classes so the complexity of the CDclasses will be sum of the complexity of all the classes; i.e.,

   CDclass = 511.1

Accordingly, the value of MCM is computed as:

$$\begin{aligned} MCM &= CIclass + CDclass + Cprocedural \\ &= 0 + 511.1 + 9 \\ &= \mathbf{520.1} \end{aligned}$$

The function point of this project is computed with the help of the count total computed in Table 4 and the value adjustment factors (VAF) based on the responses to the questions [30] given in Table 5.

Table 4

Chat Application – FP

Information Domain Value	**Weighting factor**				
	Count	**Simple**	**Average**	**Complex**	**Total**
EIs	17	3	**4**	6	68
EOs	78	**4**	5	7	312
EQs	1	3	4	**6**	6
ILFs	0	7	10	15	0
EIFs	17	5	**7**	10	119
Count Total	[redacted]	[redacted]	[redacted]	[redacted]	505

Table 5

Responses of questions for VAF

FP = count total* [0.65 + 0.01 x ∑ **(Fi)**]

FP Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**∑(Fi)**
Value Adjustment factor	0	5	5	3	1	3	3	0	3	2	3	0	0	4	**32**

= 505 x [0.65 + 0.01 x 32]

= **489.85**

Once we compute the MCM and function point (FP), we finally have to compute the code quality of the project.

The code complexity of this project is computed as,

CQ = (FP / MCM) * 10000

CQ = (489.85 / 520.1) * 10000

CQ = **9418.38108**

The code quality of this project is computed as 9418, which represents that the complexity of this project is not very high. In fact, CQ values are inversely proportional to complexity, i.e. high CQ values correlates to low complexity. This point will be clearer when we compare this project with a more complex project presented and computed in the next section.

4.2 Microprocessor Simulator

Our second project is a Microprocessor simulator. This is a simple 8085 simulator program developed in JAVA [35]. This project includes 16 classes and encompasses numerous nested loops. Its number of LOC is 2332. Due to its extremely complex structure and simpler functionalities, it has a lower CQ value.

We have to compute MCM and the function points to measure the code quality of this project. Table 6 shows the different parameters of MCM for all the classes of the project. Based on the complexity values, we have devised a graph for the complexities of all the classes in Figure 2.

Table 6

Microprocessor Simulator – Classes

class	att	str	var	obj	MA	AM	cohesion	Comp.
UserRam	1	14	0	8	2	1	2	21
RunPro	1	5	0	0	2	1	2	4
Proceed1	22	2512	15	0	7	22	0.3	2548.7
Proceed	17	5454	25	2	13	17	0.7	5497.3
SetFlag	5	44	0	0	1	5	0.2	48.8
RunErrors	0	12	0	8	0	0	0	20
MemArea	4	18	4	6	2	4	0.5	31.5
InstArea	2	18	0	24	1	1	1	43
SetC	2	15	3	0	1	2	0.5	19.5
Check	1	22	8	0	1	1	1	30
Check1	2	50	0	0	1	2	0.5	51.5
About	0	10	0	12	0	0	0	22
Check2	4	16	2	0	1	4	0.2	21.8
Check3	2	18	2	0	1	2	0.5	21.5
Check4	3	27	2	0	1	3	0.3	31.7
FlagsWindow	0	10	2	8	0	0	0	20

The graph in Figure 2 reflects the trends of the complexity of the classes of the projects. The average complexity of the classes is 527, which shows that, in general, the complexities of the classes are high. The highest complexity is 5497, which belongs to the class Proceed and is a consequence of the fact that this class contains the highest number of strings and that the complexity created by these strings is 5454. The lowest complexity belongs to the class RunPro, which is 4 due to its lowest amount of strings and attributes.

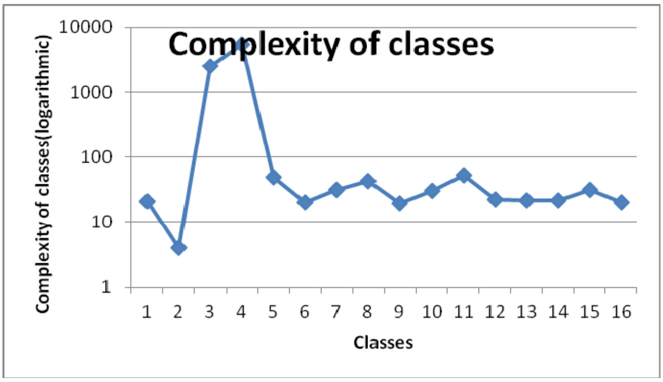


Figure 2

Complexity of Various Classes of the Microprocessor Simulator

We have to compute the complexity of the main program to calculate the procedural complexity (Cprocedural) of the project. The main program of the project has no variables, strings or objects, so complexity of this part is estimated as zero (Table 7).

Table 7

Microprocessor Simulator – Cprocedural

Non-Class	var+str+obj	Complexity
Cprocedural	0	0

Additionally, this project has several inheritance hierarchies. Figure 3 demonstrates the hierarchies among different classes. In fact, these hierarchies are the main reason of the increment of the overall complexity of the project.

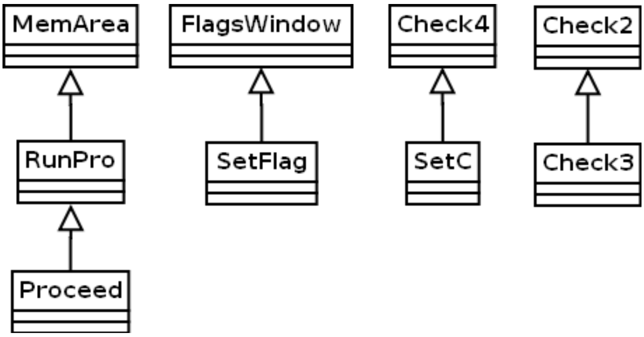


Figure 3

Microprocessor – Inheritance

Figure 3 shows that five classes are the child/subclasses in four different hierarchies. In one of the hierarchies, the depth of the inheritance tree is two.

Proceed class is at level two in the hierarchy. In the formulation of the complexity of the classes due to inheritance (CIclass), the complexities of classes are multiplied by each other. Accordingly, the total complexity of the classes caused by inheritance is computed as:

CIclass =Complexity of (MemArea*(RunPro*(Proceed))+ FlagWindows*SetFlag + Check4*SetC + Check2*Check3)

=31.5(4(5497.3)) + 20(48.8) + 31.7(19.5) + 21.8(21.5)

= 692659.8 + 976 + 618.15 + 468.7

= 694722.7

The classes which are not in the hierarchy are treated as distinct classes. The total complexities of the distinct classes are computed as:

CDclass = 2736.2

Subsequently, the complexity of overall projects is calculated as:

MCM = CIclass + CDclass + Cprocedural

= 694722.7 + 2736.2+0

= **697458.9**

The function point calculation is estimated with the information domain values for the project (to compute the total count) and value adjustment factors, as given in Tables 8 and Table 9, respectively.

Table 8

Microprocessor Simulator – FP

Information Domain Value	**Weighting factor**				
	Count	**Simple**	**Average**	**Complex**	**Total**
EIs	0	3	4	6	0
EOs	17	4	**5**	7	85
EQs	0	3	4	6	0
ILFs	0	7	10	15	0
EIFs	14	5	**7**	10	98
Count Total					183

Table 9

Responses of questions for VAF

FP = count total* [0.65 + 0.01 x **∑(Fi)**] = 183 x [0.65 + 0.01 x 19]

FP Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**∑(Fi)**
Value Adjustment factor	0	3	0	2	0	0	0	0	3	2	5	0	0	4	**19**

= **153.72**

The Code quality of the project is computed as:

CQ = (FP / MCM) * 10000

CQ = (153.72 / 697458.9) * 10000

CQ = **2.20400**

The code quality of this project is estimated as 2.20400.

Now we compare the above two projects. The code quality of both projects are computed as 9418 and 2.20. According to the structure of the metric, a high value of CQ represents low complexity and vice-versa. This means that the second project is comparatively more complex than the first project because its CQ value of 2.20 is much smaller than the 9418. The number of classes in the chatting application and Microprocessor simulator are 13 and 16, respectively, which are not too different (in terms of number). However, the average complexity of the Microprocessor class is 527 and the Chat application is 39, which reflects that the complexity of classes in Microprocessor simulator is much more complex in comparison to the classes of chatting applications. The complexity of the Microprocessor simulator is high because it contains a complex structure characterized by several nested loops.

The above two projects are different in nature. The MCM has well differentiated both projects in terms of their complexities. These experimentations prove the applicability of multi-paradigm metric in JAVA projects.

Conclusion and Future Work

A multi-paradigm complexity metric is evaluated through measurement theory and applied to the two JAVA projects. The evaluation of compliance with measurement theory has proved that this metric satisfies the additive property. This additive nature of the metric proves its theoretical soundness. Furthermore, the metric is applied to two real JAVA projects. The projects are different in nature (in terms of their architecture of the classes), and the MCM demonstrates a good differentiation between them in terms of their complexity, which reinforces that the MCM is useful in estimating the complexity of JAVA projects. As a future work, we aim to fix the thresholds [36] for MCM. To achieve thresholds, we will perform more experimentation on real projects in industry. We also plan to apply the MCM on projects developed in other languages.

References

[1] Francalanci, C., Merlo, F.: The Impact of Complexity on Software Design Quality and Costs: An Exploratory Empirical Analysis of Open Source Applications White paper available at: (last accessed 16.03.2010) http://is2.lse.ac.uk/asp/aspecis/20080122.pdf

[2] Banker R. D., Datar S. M., Zweig, D. (1989) Software Complexity and Maintainability, Proceedings of the tenth international conference on Information Systems, pp. 247-255, Boston, Massachusetts, United States

[3] Dawei E. (2007) The Software Complexity Model and Metrics for Object-Oriented, In Proc. of IEEE International Workshop on Anti-counterfeiting, Security, Identification, pp. 464-469

[4] Dufour B., Driesen K., Hendren L., Verbrugge C. (2003) Dynamic Metrics For JAVA, In Proceedings of the Conference On Object-Oriented Programming, Systems, Languages, and Applications, October 26-30, 2003, Anaheim, California, USA, pp. 149-168

[5] Cahoon B., McKinley K. S. (2001) Data Flow Analysis for Software Prefetching Linked Data Structures In JAVA Controller. In Proc. of The 2001 International Conference on Parallel Architectures and Compilation Techniques, pp. 280-291, September 2001, Barcelona, Spain

[6] Sundaresan V., Hendren L., Razafimahefa C., Rai R. V., Lam P., Gagnon E., Godin C. (2000) Practical Virtual Method Call Resolution for JAVA. In Proceedings of the Conference on Object-Oriented Programming, Systems, Languages, And Applications, pp. 264-280, ACM Press

[7] Vijaykrishnan N., Ranganathan N. (1999) Tuning Branch Predictors to Support Virtual Method Invocation In JAVA. In Proceedings of the 5th USENIX Conference on Object-Oriented Technologies and Systems, May 1999, pp. 16-16

[8] Qian F., Hendren L., Verbrugge C. (2002) A Comprehensive Approach To Array Bounds Check Elimination For JAVA. In Proc. of the International Conference on Compiler Construction, Lecturer Notes in Computer Science, 2304, pp. 325-341

[9] Ruf E. (2000) Effective Synchronization Removal for JAVA. In Proc. of the ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 208-218

[10] Shuf Y., Serrano M. J., Gupta M., Singh J. P. (2001) Characterizing the Memory Behavior of JAVA Workloads: A Structured View and Opportunities for Optimizations. In Proceedings of the 2001 ACM SIGMETRICS International Conference on Measurement and Modelling of Computer Systems, pp. 194-205

[11] Mäkelä S. Leppänen V. (2009) Client-based Cohesion Metrics for JAVA Programs, Science of Computer Programming, 74(5/6), pp. 355-378

[12] Nasseri E. Counsell S., Shepperd M. (2010) Class Movement and Re-Location: An Empirical Study of JAVA Inheritance Evolution, Journal of Systems and Software, 83(2) pp. 303-315

[13] Kaczmarek J. Kucharski M. (2004) Size and Effort Estimation for Applications Written in JAVA, Information and Software Technology, 46(9) pp. 589-601

[14] Pirró G. (2009) A Semantic Similarity Metric Combining Features and Intrinsic Information Content, Data & Knowledge Engineering, 68(11) pp. 1289-1308

[15] Romano D. (2011) Using Source Code Metrics to Predict Change-Prone JAVA Interfaces, In Proc of 27th IEEE International Conference on Software maintenance, pp. 303-312

[16] Sommerville, I. (2004) Software Engineering, 7th Edition, Addison Wesley, 2004

[17] Misra S., Akman I., Cafer F. (2011) A Multi Paradigm Complexity Metric, Lecture Notes in Computer Science, 6786/2011, pp. 342-354

[18] Albrecht A. J. (1979) Measuring Application Development Productivity, Proceedings of the Joint SHARE, GUIDE, and IBM Application Development Symposium, Monterey, California, October 14-17, IBM Corporation (1979) pp. 83-92

[19] Misra S, Cafer F. (2011) Estimating Complexity of Programs in Python Language Technical Gazette, 18 (1) pp. 1-10

[20] Misra S. (2011) An Approach for Empirical Validation Process for Software Complexity Measures, Acta Polytechnica Hungarica, 8(2), pp. 141-160

[21] Misra S., Akman I. (2010) Unified Complexity Metric: A Measure of Complexity, Proc. of National Academy of Sciences Section A. 80(2) pp. 167-176

[22] Misra S., and Akman I. (2008) Weighted Class Complexity: A Measure of Complexity for Object Oriented Systems, Journal of Information Science and Engineering, 24, pp. 1689-1708

[23] Misra S., Akman I., Koyuncu M. (2011) An Inheritance Complexity Metric for Object Oriented Code: A Cognitive Approach, SADHANA (Springer), 36(3) pp. 317-338

[24] Basci D., Misra S. (2011) Metrics Suite for Maintainability of XML Web-Services' IET Software 5(3), pp. 320-341

[25] Basci D., Misra S. (2009) Data Complexity Metrics for Web-Services, Advances in Electrical and Computer Engineering, 9(2), pp. 9-15

[26] Basci D., Misra S. (2011) Entropy as a Measure of Complexity of XML Schema Documents' Int. A. journal of Information Technology, 8(1) pp. 16-25

[27] Basci D., Misra S. (2009) Measuring and Evaluating a Design Complexity Metric for XML Schema Documents, Journal of Information Science and Engineering, 25(5) pp. 1405-1425

[28] Tonbul G. and Misra S. (2009) Error Density Metrics for Business Process Modeling, In Proc. of the 24th International Symposium on Computer and Information Sciences, pp. 542-546

[29] Lara, P., Fernandez, l. (2008) Test Case Generation, UML and Eclipse, Dr.Dobbs Journal, 22 (11) pp. 49-52

[30] Roger S. P. (2005) Software Engineering – A practitioner's approach, 6th Edition. McGraw-Hill

[31] Briand L. C., Morasca S., Basily V. R.(1996) Property-based Software Engineering Measurement, IEEE Transactions on Software Engineering, 22 (1), pp. 68-86

[32] Gupta V, Chhabra J. K. (2009) Package Coupling Measurement in Object-oriented Software. Journal of Computer Science and Technology 24(2), pp. 273-283

[33] Costagliola G., Ferrucci F., Tortora G., Vitiello G. (2005) Class Points: An Approach for the Size Estimation of Object-Oriented Systems, IEEE Transactions on Software Engineering, 31(1) pp. 52-74

[34] Source Codes World – Chatting Application (last accessed 21.02.2010)
Available at: http://www.sourcecodesworld.com/source/show.asp?ScriptID=524

[35] Source Codes World – Microprocessor Simulator (last accessed 21.02.2010)
Available at: http://www.sourcecodesworld.com/source/show.asp?ScriptID=849

[36] Misra S. (2011) Evaluation Criteria for Object-oriented Metrics, Acta Polytechnica Hungarica, 8(5), pp. 109-136

Hydrodynamic Optimization of Marine Propeller Using Gradient and Non-Gradient-based Algorithms

Ramin Taheri, Karim Mazaheri

Laboratory of Hydrodynamic Computation and Optimization Design, Center of excellence in Aerospace Engineering Systems, Sharif University of Technology, Azadi Ave., Tehran, Iran, POBox:11365-8629
E-mail: ramintaheri@ae.sharif.ir; mazaheri@sharif.edu

Abstract: Here a propeller design method based on a vortex lattice algorithm is developed, and two gradient-based and non-gradient-based optimization algorithms are implemented to optimize the shape and efficiency of two propellers. For the analysis of the hydrodynamic performance parameters, a vortex lattice method was used by implementing a computer code. In the first problem, one of the Sequential Unconstraint Minimization Techniques (SUMT) is employed to minimize the torque coefficient as an objective function, while keeping the thrust coefficient constant as a constraint. Also, chord distribution is considered as a design variable, namely 11 design variables. In the second problem, a modified Genetic algorithm is used. The objective function is to maximize efficiency by considering the design variables as non-dimensional blade's chord and thickness distribution along the blade, namely 22 design variables. The hydrodynamic performance analyzer code is modified by a higher order Quasi-Newton scheme. Also, a hybrid function is used to improve the accuracy of the convergence. The solution of the optimization problems showed that a nearly 13% improvement in efficiency and a nearly 15% decrease in torque coefficient for the first propeller, as well as nearly 10% improvement for efficiency of the later propeller, is possible.

Keywords: Marine propeller; gradient-based optimization algorithm; Genetic algorithm; Vortex lattice

1 Introduction

The complexity of the flow field in which the propeller must operate efficiently will lead a designer to lay out a propeller to overcome most of the dilemma. Another difficulty which arises during propeller action is the variation of inflow, which has a great influence on propellers. Hence, the range of design is restricted for designers.

The development of the Momentum theory for marine propellers was the starting point of the aerodynamic analysis of rotary wings. Betz [1] firstly introduced the lifting line theory and Goldstein [2] and Lerbs [3] consequently improved the method. Theodorson extended the vortex theory for highly loaded propeller. Rand and Rosen [4], Chang and Sullivan [5] and Chiu and Peters [6] used the lifting line theory for their works. Later on, Eckhart and Morgan [7] proposed the Lifting-Surface correction factors that were then developed further by Pien [8] and Kerwin [9]. Chord distribution, wing tip shape and twist angle were shown by McVeigh and McHugh [10] and Walsh et al. [11] to be the main factors which control the performance of straightened blade propellers. Lee [12] applied the vortex lattice methods for the prediction of the hydrodynamic performance of marine propellers. Khot and Zweber [13] optimized the structure of a composite wing by using gradient based algorithm. The twist angle distribution and a span wise chord distribution were optimized by Cho and Lee [14] utilizing gradient based optimization with the penalty function method. Also, investigating the possibility of maximizing the efficiency by utilizing Genetic algorithm was done by Lee and Lin [15]. Later on, Plucinski et al. [16] optimized a self-twisting propeller, using a genetic algorithm by considering the orientation angles of the fibers in each layer as the design variables for efficiency improvement. For design optimization, a propeller performance analysis program was developed and integrated into a genetic algorithm by Christoph Burger [17]. The duty of the tool is to produce optimal propeller geometry for a given aim, which includes performance and/or acoustic signature. Using a genetic optimization algorithm, Aykut et al. [30] achieved a more convincible result compared to previous studies. Taheri et al. [18, 19] studied the process of both gradient and non-gradient-based optimization algorithms. Also, the accomplishment of an inverse design as well as the optimization of the propeller were studied and done by Taheri et al. [20, 21].

2 Openprop

OpenProp is a design and analysis tool for propellers and turbines. The code is written in MATLAB M-code and the numerical model is based on vortex lattice lifting line methods. The capability of the code has been tested by validating an experiment results and the numerical method which is used in OpenProp. OpenProp began in 2001 and was further developed by Kerwin [22] in 2007. The code was improved by Stubblefield in 2008 and Epps in 2009, respectively. Development of the OpenProp code suite began at MIT in 2006 under the direction of Kimball as a Matlab version of a Fortran code published by Kerwin called PVL (Kerwin 2007). Subsequent researchers have extended the functionality of the code with the most recent version implemented by Epps (Epps, Stanway and Kimball 2009). An explanation of the theory can be found in (Epps 2010) which also presents validation of the code as presented in Figure 1 below. Further reference to the code and website can be found in (Kimball and Epps, 2010). The good agreement between experimental data and numerical calculations done by OpenProp and a commercial package (Propeller Blade Design, PBD) is shown here.

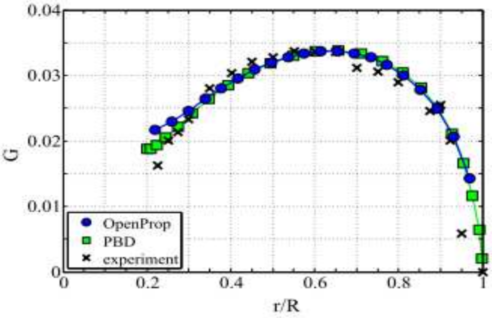


Figure 1

Validation of Circulation results calculated by OpenProp and comparison with experimental data and PBD (Propeller Blade Design), reprinted with permission from Epps [23]

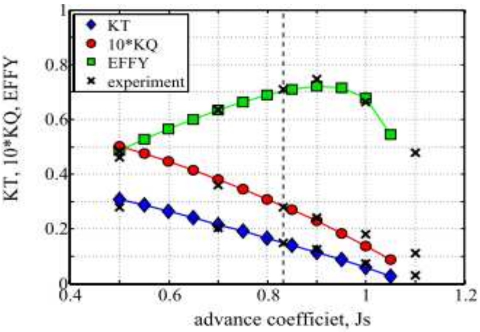


Figure 2

Validation of efficiency calculated by OpenProp and comparison with experimental data, reprinted with permission from Epps [23]

First, the circulation distribution is compared, as is shown in Figure 1. Second, thrust and torque coefficients, as well as final efficiency distribution over the range of propeller performance, has been done and the results are shown in Fig. 2. The above illustrations convinced us to rely on this code and use that as a package to calculate our hydrodynamic performance needs.

3 Propeller Lifting Line Formulation

In the following calculation, based on moderately loaded lifting line theory, the lifting line is the representative of a propeller blade, with trailing vorticity aligned to the local flow velocity (i.e., the vector sum of free-stream plus induced velocity). Using a vortex lattice with helical trailing vortex filaments shed at discrete stations along the blade, induced velocities can be computed. The blade is sectioned discretely, having 2D section properties at each radius. Loads are computed by integrating the 2D sections load over the span of the blade. The velocities and forces (per unit span) on a 2D blade section can be seen in both the axial and tangential directions in "Figure 3". Apparent tangential inflow at radius r is $-\omega r e_t$, while the propeller shaft rotates with angular velocity of ωe_a. Total resultant inflow velocity, V^*, and its orientation pitch angle can be computed by equation (1) and equation (2), respectively.

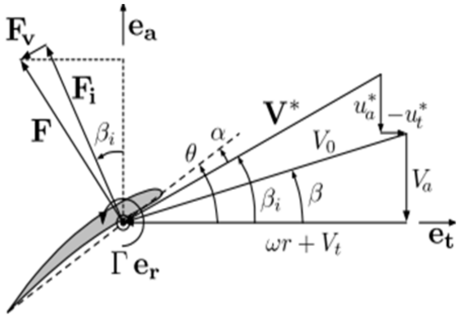


Figure 3
Propeller velocity/force diagram, as viewed from the tip towards the root of the blades. All velocities are relative to a stationary blade section at radius r, reprinted with permission from Epps [23].

This is an equation example:

$$V^* = \sqrt{(V_a + u_a^*)^2 + (\omega r + V_a + u_t^*)^2}. \tag{1}$$

$$\beta_i = \arctan(\frac{V_a + u_a^*}{\omega r + V_a + u_t^*}), \tag{2}$$

where $V_a = -V_a e_a$ and $V_t = -V_t e_t$ are the axial and tangential inflow velocities, $u_a^* = -u_a^* e_a$ and $u_t^* = -u_t^* e_t$ are induced axial and tangential velocities, α is the angle of attack, $\theta = \alpha + \beta_i$ is blade pitch angle, Γe_r is circulation, $F_i = \rho V^* (\Gamma e_r)$ is (inviscid) Kutta-Joukowski lift force, and F_v is viscous drag force aligned with V^*. Assuming the Z blades are identical, the total thrust and torque on the propeller are

$$T = z\int_{r_h}^{R} (F_i \cos\beta_i - F_v \sin\beta_i) dr(\hat{e_a}), \tag{3}$$

$$Q = z\int_{r_h}^{R} (F_i \sin\beta_i + F_v \cos\beta_i) rdr(-\hat{e_a}), \tag{4}$$

where $F_i = \rho V^* \Gamma$ and $F_v = \frac{1}{2}\rho V^{*2}(C_D)c$ are the magnitude of inviscid and viscous force per unit radius, ρ is the fluid density, C_D is the section drag coefficient, c is the section chord, and r_h and R are the radius of the hub and blade tip, respectively.

The propeller power consumption is the product of torque and angular velocity

$$P = Q\omega, \tag{5}$$

The propeller puts power into the fluid when, $P \succ 0$ (i.e. the torque resists the motion). As the useful power produced by the propeller is TV_s, where V_s is the ship speed (i.e. free stream velocity), the efficiency of propeller is defined by [21]

$$\eta = \frac{TV_s}{Q\omega}. \tag{6}$$

After the above calculations, thrust and torque coefficients as well as advanced ratio are calculated as follow

$$K_T = \frac{T}{\rho n^2 D^4}. \tag{7}$$

$$K_Q = \frac{Q}{\rho n^2 D^5}. \tag{8}$$

$$J = \frac{V_s}{nD}. \tag{9}$$

4 Inverse Design

First, to show the capability of the hydrodynamic analyzer code, the inverse design was done by a nearly ill-posed initial guess. This part was only done to prove the validity of the results that had been calculated by OpenProp code. For this purpose, the circulation distribution along the blade was chosen to reach our desired circulation distribution. Following is the function with which we explored the validation of the code

$$I = \int \left| G - G_{desired} \right|. \tag{10}$$

The result of this calculation is shown in "Figure 4".

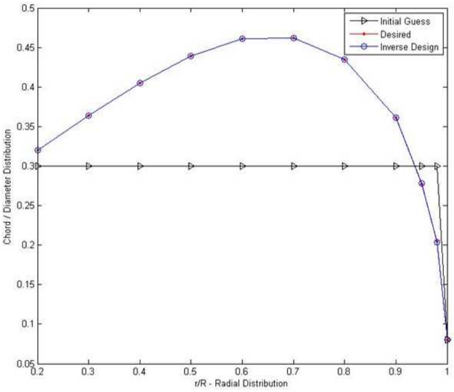


Figure 4
Results of propeller inverse design

4 Optimization Algorithms

Generally speaking, optimization algorithms are categorized as two major sets. The first one is gradient-based algorithms and the other one is non-gradient-based algorithms. The act of choosing each set depends on the pros and cons related to each category, as well as to the specific conditions for a problem.

4.1 Gradient-based Algorithm

For Gradient-based algorithms, the simplest way of calculating derivatives for functions is the Finite Difference method. If the round-off error is important, the Complex method will be beneficial. There are many methods to calculate the derivatives, such as sequential quadratic programming (SQP), which is the most popular method for constraint optimization problems. For unconstraint optimization problems, Quasi-Newton methods play an important role. Also, by adding penalty terms to constraint formulation one can obtain an unconstraint approach. One of the sequential unconstraint optimization techniques, called the extended linear interior penalty function method, (EIPM) was used here. One of the most considerable differences between gradient-based and non-gradient-based algorithms is the existence of linear trend associated with number of design variables, while in the latter methods the cost of calculations is increased drastically by an increase in number of design variables.

The Extended linear Interior Penalty function Method (EIPM), one of the sequential Unconstrained Minimization Techniques (SUMT), is employed as one of the optimization techniques. The aforementioned technique transforms a constrained optimization problem into a series of unconstrained optimization problems and constructs a pseudo-objective function using penalty functions.

A constrained optimization problem is stated as [24]:

$$\begin{aligned}&\text{minimize}\, f(x)\\&\text{subject to}\; g_j(x) \le 0,\ \text{j=1,m}\\&\mathrm{h}_k(x) = 0,\ k = 1, l\end{aligned} \tag{11}$$

where $f(x)$ is objective function. $g_j(x)$ and $h_k(x)$ are inequality and equality constraints, respectively. The transformed unconstrained optimization problem is also stated as:

$$pseudo - objective\ \phi(x, r_p, \dot{r}_p) = f(x) + r_p P(x) \tag{12}$$

where r_p is a multiplier and will increase in each iteration until it reaches to a pre-defined value, and $P(x)$ is a penalty function consists of equality and inequality constraints. The final form of the transformed constrained optimization problem is

$$S^k = -H^k \nabla F(x^k) \tag{13}$$

$$H^{k+1} = H^k + D^k \tag{14}$$

where D^k is defined as follows [25]:

$$\left[D^k\right] = \left[M_i\right] + \left[N_i\right] \tag{15}$$

$$\left[M_i\right] = \lambda_i^* \frac{S_i S_i^T}{S_i^T g_i} \tag{16}$$

$$\left[N_i\right] = -\frac{\left(\left[H_i\right] g_i\right)\left(\left[H_i\right] g_i\right)^T}{g_i^T \left[H_i\right] g_i} \tag{17}$$

where $g_i = \nabla f\left(X_{i+1}\right) - \nabla f\left(X_i\right) = \nabla f_{i+1} - \nabla f_i$.

The iterative procedure is,

I. Start with initial design variables as a vector x^0 and initiate Hessian matrix H^0. (commonly identity matrix, $H^0 = I$)

II. Finding search direction, $S^k = -H^k \nabla F(x^k)$

III. Finding step length, λ by a univariate optimization process which determines the amount of change in the search direction.

IV. Updating design variables and then calculating gradient and Hessian of the function,

$$x^{k+1} = x^k + \lambda S^k$$

V. Checking convergence criteria and going to step II.

Here, non-dimensional chord distribution is considered as design variables, in fact 11 variables.

The test case which has been used was a DTMB 4119 propeller of which the geometry characteristics are listed in "Table 1".

Table 1
Geometry definition of DTMB 4119 propeller [26]

r / R	c / D	p / D	qr	IT / D	tm / C	fm / C
0.2	0.32	1.105	0	0	0.2055	0.01429
0.3	0.3635	1.102	0	0	0.1553	0.02318
0.4	0.4048	1.098	0	0	0.1180	0.02303
0.5	0.4392	1.093	0	0	0.09016	0.02182
0.6	0.4610	1.088	0	0	0.0696	0.02072
0.7	0.4622	1.084	0	0	0.05814	0.02003
0.8	0.4347	1.081	0	0	0.04206	0.01967
0.9	0.3613	1.079	0	0	0.03321	0.01817
0.95	0.2775	1.077	0	0	0.03228	0.01631
1.0	0.0	1.075	0	0	0.0316	0.01175

Other characteristics that should be considered are:

1. The propeller inflow is uniform.
2. The propeller has 3 blades, i.e. N = 3.
3. The hub-to-diameter ratio is 0.2.
4. The propeller has no skew and no rake.
5. The blade sections are designed with NACA 66 modified profiles and a camber line of a = 0.8.
6. The propeller advanced ratio is $j = 0.833$.
7. The direction of rotation is right-handed.

4.1.1 Objective Function

For the abovementioned gradient-based optimization algorithm, the objective function would be the torque, and the equality constraint function would be given in terms of thrust as below:

$$\begin{aligned} &Minimize\, f(x) \;=\; \mathrm{k}_Q(x) \\ &Subject\ to\ h(x) \;=\; \frac{k_t - k_{t0}}{k_{t0}} \end{aligned} \tag{18}$$

Where X represents design variables, f(x) is objective function, and h(x) is equality constraint.

4.2 Genetic Algorithm

Mitchel [27] states that "Genetic Algorithms were invented by John Holland in the 1960s and were developed by Holland and his colleagues at the University of Michigan in the 1960s and 1970s". Genetic algorithms are a subset of stochastic optimization methods, methods in which statistical data play an important role [28]. Literally speaking, genetic algorithms are specifically modeled after the process of natural selection. In the natural selection process it is obvious that the fittest individuals, which have a higher probability for regeneration, also have a higher probability to be chosen and being sent for the next generation.

In an optimization problem, design variables are put into chromosomes and then they undergo every chromosomal operation, such as cross-over, the process by which chromosomes exchange genes in real-world genetics, mutation and so on.

The creation of the initial population of random individuals is what the genetic algorithm begins with, each of which represents a probable solution in term of parameters. The way of well-satisfying the objective function is called fitness, which is assigned to each individual consisting of a particular set of design variables. In GA terminology, the objective function is the function that determines the performance of a particular chromosome (i.e., member of the population).

The fitter the individual, the higher chance of it having children and reproducing for the next population. This should be certified by genetic algorithm for modeling the process of natural selection efficiently. Pairing off individuals for mating and having children will occur for subsequent populations.

The simplicity of making the chance of having children proportional to the relative fitness of an individuals, and mating individuals which are competing in a domain of diverse fitness, in which higher fitness means a higher probability of winning, both can be equal.

As a matter of fact, the existence of the probability which says that a child may have a mutation in its genetic code, as well as the randomness of chromosomes which are received by father or mother, prevents the algorithm from being trapped in a globally non-optimal local minimum or maximum. Before reaching the determined number of generations or some other stopping criteria, the process of mating, mutation and cross-over will be repeated. After that, by comparing fittest individuals iteratively, the global optimum will be found. At the same time, GA has some disadvantages. As has been mentioned, in using GA there is a greater likelihood that a global optimum solution will be found. However, finding this global optimum is not guaranteed. Even if GA is in the neighborhood of the global optimum, there is the possibility that through crossover and mutation the global optimum may not be achieved. Also, GA does not address the robustness of the individual design solutions it creates. GA simply attempts to meet the desired goals and will adjust the design parameters accordingly. Thus, it is up to the user to ensure the proper operation of GA and to verify that the results are genuine. Finally, the satisfactory operation of GA relies on the accuracy of the system models that make up the objective function.

Here, in order to create the genetic algorithm being modified, a hybrid function is used. The hybrid function enables us to specify another minimization function that runs after the genetic algorithm terminates. Also, the hybrid function is a function by which we can cause the genetic algorithm to be converged faster. For instance, after reaching to the region of a highly likely fitter population for the next generation, the hybrid function, by using a simple gradient-based algorithm (Steepest Descent), helps us to put the design variables in a more efficient direction towards an optimum point, instead of just producing a large number of populations. Hence, using this scheme decreases the time needed for convergence. To implement this method, a gradient-based algorithm is used to lower the computational costs. Since in some manners of using genetic algorithm it is obvious to find better population for the next generation near the solution point, using a gradient-based algorithm can be considered very helpful and reliable.

In the GA algorithm, the population type specifies the type of input to the fitness function. Types and their restrictions here is a double vector, and the population size is 20.

A scaling function is the function that converts raw fitness scores returned by the fitness function to values in a range that is suitable for the selection function. Here we have used ranked scaling, which scales the raw scores based on the rank of each individual, rather than its score. The rank of an individual is its position in the sorted scores. The rank of the fittest individual is 1, the next fittest is 2, and so on. Rank fitness scaling removes the effect of the spread of the raw scores.

The selection function chooses parents for the next generation based on their scaled values from the fitness scaling function. Here we have used Stochastic Uniform, which lays out a line in which each parent corresponds to a section of the line of length proportional to its expectation. The algorithm moves along the line in steps of equal size, one step for each parent. At each step, the algorithm allocates a parent from the section it lands on. The first step is a uniform random number less than the step size.

Crossover combines two individuals, or parents, to form a new individual, or child, for the next generation. Here we have used Scattered, which creates a random binary vector. It then selects the genes where the vector is a 1 from the first parent, and the genes where the vector is a 0 from the second parent, and combines the genes to form the child.

4.2.1 Objective Function

$$\begin{aligned}&\text{minimize}\, f(x)\\&\text{subject to}\; g_j(x) \le 0,\; \text{j=1,m}\\&\text{h}_k(x) = 0,\; k = 1, l\end{aligned} \tag{19}$$

where h_k and g_j are equality and inequality constraints, respectively. Herein, thickness distribution and non-dimensional chord distribution are considered as design variables, in fact 22 design variables, under the following conditions as an inequality constraint:

- For structural requirements, minimum foil section thickness should be considered.
- To avoid cavitation, the minimum pressure coefficient should be negative and the cavitation is also depth dependent.

The test case used was a DTRC 4119 propeller, the geometry characteristics of which are listed in Table 2.

Table 2

Geometry definition of DTRC 4119 propeller [26]

r/R	c/D	p/D	qr	IT/D	tm/C	fm/C
0.2	0.32	1.105	0	0	0.2055	0.01429
0.3	0.3635	1.102	0	0	0.1553	0.02318
0.4	0.4048	1.098	0	0	0.1180	0.02303
0.5	0.4392	1.093	0	0	0.09016	0.02182
0.6	0.4610	1.088	0	0	0.0696	0.02072
0.7	0.4622	1.084	0	0	0.05814	0.02003
0.8	0.4347	1.081	0	0	0.04206	0.01967
0.9	0.3613	1.079	0	0	0.03321	0.01817
0.95	0.2775	1.077	0	0	0.03228	0.01631
1.0	0.0	1.075	0	0	0.0316	0.01175

where r/R is non-dimensional radius distribution, c/D is non-dimensional chord distribution, p/D is non-dimensional pitch distribution, qr is the rake of propeller, IT/D is the non-dimensional skew of propeller, tm/C is the non-dimensional maximum thickness distribution, and fm/C is the non-dimensional camber distribution.

Other characteristics that should be considered are:

1. The propeller inflow is uniform.
2. The propeller has 3 blades, i.e. N = 3.
3. The hub-to-tip diameter ratio is 0.2.
4. The propeller has no skew and no rake.
5. The blade sections are designed with NACA 66 modified profiles and a camber line of a = 0.8.
6. The propeller advanced ratio is $J = 0.833$.
7. The direction of rotation is right-handed.

The whole procedure is depicted in Figure 5. First, the code is started from an initial guess which is our so-called propeller. Then, Openprop is used to calculate the hydrodynamic parameters that are compulsory for the objective function. The next step is to find better geometry by implementing the modified genetic algorithm. Finally, checking the convergence criteria is done iteratively to quit the program.

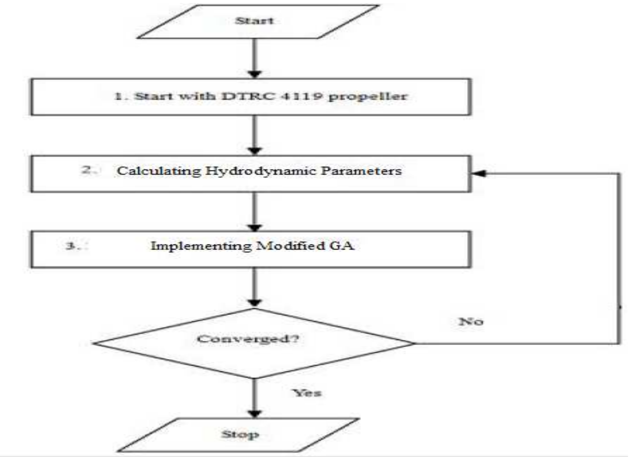

Start
1. Start with DTRC 4119 propeller
2. Calculating Hydrodynamic Parameters
3. Implementing Modified GA
Converged?
No
Yes
Stop

Figure 5
Flowchart for optimization process

5 Results and Discussion

The results for the circulation distribution as well as the torque coefficient for the first optimization problem are as follows, where higher circulation keeps the efficiency higher (Figure 6 and Table 3):

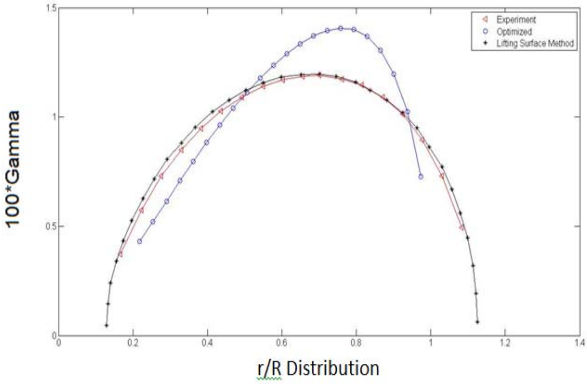


Figure 6

Radial Circulation distribution. Initial (Lifting Surface Method), Experiment, and Optimized

Table 3

Optimized propeller in comparison to the original one

Type of propeller	K_t	K_Q	η
DTMB 4119	0.1468	0.0264	0.7375
Present (Optimized)	0.147	0.0227	0.8589

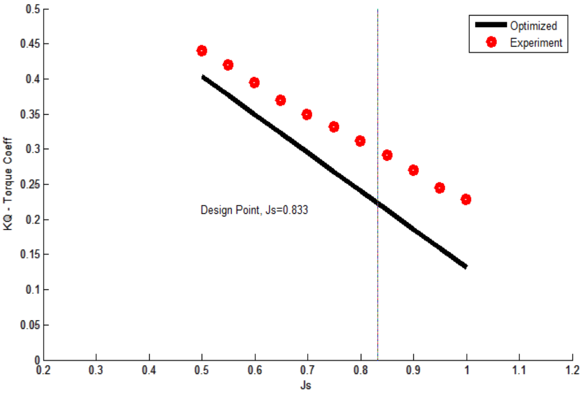


Figure 7

Reduction in torque coefficient distribution with respect to advanced ratio. Initial (Lifting Surface Method), and Optimized.

For the second problem, the combination of both hydrodynamic analyzer code and modified Genetic algorithm has been done elaborately and the results of any improvement are shown in the next consequent figures. The shape of the cross sectional airfoils is depicted in Figure 7. Furthermore, the elongated chord distribution for optimized propeller in comparison with initial one can be seen in Figure 9. A little ripple at the top most of blade's tip is also demonstrated, which is similar to the behavior which was shown by Karim [29]. The time needed for calculation was approximately 10 minutes for 113 generations on a personal computer with a CPU speed of 3.1 GHz. The history of the convergence is shown in Figure 10. Scrutinizing Figure 11, it is obvious that efficiency at the predetermined advanced ratio is higher than those for which calculation was done for the initial and redesigned ones [29]. In order to reach higher efficiency, as shown in both Figure 12 and Figure 13, the thinner camber distribution, whether non-dimensional distribution or just the very distribution, should accommodate with the length of the blade. Although the new design for the propeller has a lower thrust coefficient, it is clear that this penalty can be compensated for by higher efficiency, by which the performance of propeller are put considered. Table 3 shows the thrust and torque coefficients as well as efficiency over time. It is very clear that this efficiency is high enough to convince a propeller designer not only of the reliability of the method, but also the usefulness of the tool.

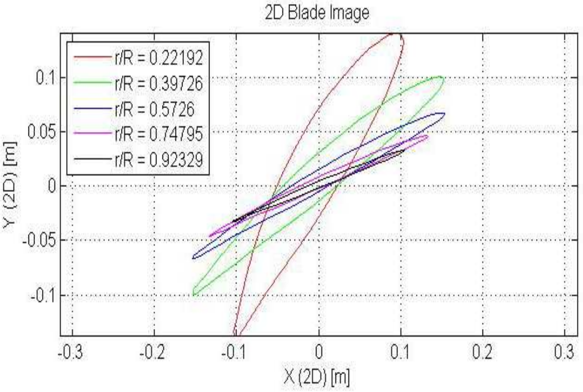


Figure 8

Two dimensional cross section distribution for optimized propeller

Conclusions

In this article, to validate our scheme, first we note that our physical analysis tool (OpenProp) is clearly validated, as shown in Figure 1, Figure 2 and "Figure 4" for our propellers. The result of OpenProp has been compared with experimental data.

Here we applied two optimization algorithms to a given propeller geometry. For gradient based algorithm we have used non-dimensional chord distribution as design variables and have kept thrust coefficient constant as an equality constraint. Moreover, we have applied an inverse design scheme to an initial guess for our propeller geometry. Figure 4 shows, as expected, very good agreement between our results and experimental data. Table 3 compares hydrodynamic shape parameters achieved by optimization algorithms with the initial values. The efficiency improvement (nearly 13%) shows that the optimized circulation distribution was higher than the experimental values; higher circulation leads to a higher lift force and, consequently, higher efficiency can be achieved. Also, the torque coefficient reduction shown in Figure 7 demonstrates that a nearly 15% improvement can be considered possible.

The second optimization problem presented results from the application of a modified genetic algorithm technique to the design optimization of marine propeller incorporating vortex lattice method (VLM). In this research, the hybrid function was used in order to modify the genetic algorithm modified and to reach the final solution more quickly than usual. The performance of the modified genetic algorithm coupled with hydrodynamic performance analyzer code seemed to be better than one which was used by Karim [29]. Chord distribution and non-dimensional thickness distribution at twenty two sections have been optimized as design variables for efficiency improvement. Achieving a lower torque coefficient as well as higher efficiency were clearly possible. Another improvement has been applied to the hydrodynamic performance analyzer code, that is, OpenProp, to increase the accuracy of the results one order of magnitude higher than the previously best one by setting a Quasi-Newton method instead of Newton method to reach to the abovementioned purpose.

References

[1] Betz, A., “SchraubenPropeller mit geringstem Energieverlust”. K. Ges. Wiss. Gottingen Nachr. Math. Phys. Klasse, 1919

[2] Goldstein, S., “On the Vortex Theory of Screw Propellers”. Proc. R. Soc. London Ser. A 123 :440-465, 1929

[3] Lerbs, H. W., “Moderately Loaded Propellers with a Finite Number of Blades and an Arbitrary Distribution of Circulation”. SNAME Trans, 60, 1952

[4] Rand, A., and Rosen, A., “A Lifting Line Theory for Curved Helicopter Blades in Hovering and Axial Flight”. The 8th European Rotorcraft Forum, Aix-en Provence, France, 1982

[5] Chang, L. K., and Sullivan, J. P., “Optimization of Propeller Blade Twist by an Analytical Method”. AIAA Journal, 22(2), 22, 1982

[6] Chiu, Y. D., and Peters, D. A., “Numerical Solution of Induced Velocities by Semi-Infinite Tip Vortex Lines”. Journal of Aircraft, 25(8), 684, 1987

[7] Eckhart, M. K., and Morgan, W. B., “A Propeller Design Method”. SNAME Trans. 1955, 63, 1955

[8] Pein, P. C., “The Calculation of Marine Propellers Based on Lifting Surface Theory”. J. of Ship Research, Vol. 5, No. 2

[9] Kerwin, J. E., “The Solution of Propeller Lifting Surface Problems by Vortex Lattice Methods”. Report, Dep. of Ocean Eng., MIT, 2001

[10] McVeigh, M. A., and McHugh, F. J., “Influence of Tip Shape Chord, Blade Number and Airfoil on Advanced Rotor Performance”. The 38th Annual Forum of the American Helicopter Society, Anaheim, CA, 1982

[11] Walsh, J. L., and Bingham, G. J., and Riley, M. F., “Optimization Method Applied to the Aerodynamic Design of Helicopter Rotor Blades”. The 26th AIAA/ASME/ASCE/AHS structures, Structural Dynamic and Material Conference, Orlando, Florida, 1985

[12] Lee, C. S., Prediction of Steady and Unsteady Performance of Marine Propellers with or without Cavitation by Numerical Lifting Surface Theory. Ph. D. Thesis, MIT., USA, 1979

[13] Khot, N. S., and Zweber, J. V., “Design of Flutter Characteristics of Composite Wings Using Frequency Constraint Optimization”. Journal of Aerospace Engineering, Vol. 16, pp 19-30, 2003

[14] Cho, J., Lee, S-C., “Propeller Blade Shape Optimization for Efficiency Improvement”. Computers and Fluids, Vol. 27, No. 3, pp. 407-419, 1998

[15] Lee, Y-J., and Lin, C-C., “Optimized Design of Composite Propeller”. Mechanics of Advanced Materials and Structures. Vol. 11, pp. 17-30, 2004

[16] Plucinski, M. M., and Young, Y. L., and Liu, Z., “Optimization of a Self-Twisting Composite Marine Propeller Using Genetic Algorithms”. 16th International Conference on Composite Materials, Kyoto, Japan, 2007

[17] Burger, C., “Propeller Performance Analysis and Multidisciplinary Optimization Using a Genetic Algorithm”. ProQuest Publisher, PHD Thesis, Auburn University, 2007

[18] Taheri, R., Mazaheri, k., “*Comparison of Gradiend-based and Genetic Optimization Algorithms Applied to wing-body Configuration*”. 20th Annual International Conference on Mechanical Engineering-ISME2012, School of Mechanical Eng., Shiraz University, Shiraz, Iran, 16-18 May, 2012

[19] Taheri, R., Mazaheri, k., “*Rapid Aerodynamic Optimization of Wing and Body ConfigurationUsing Gradient Based Approach* ”. 11th Annual International Conference on Aerospace Engineering, Shahid Sattary University of Science and Technologies, Tehran, Iran, 2-4 March, 2012

[20] Taheri, R., Mazaheri, K., "*Blade Shape Optimization of Marine Propeller via Genetic Algorithm for Efficiency Improvement*". Proceedings of ASME Turbo Expo 2012, Copenhagen, Denmark, June 11-15, 2012

[21] Taheri, R., Mazaheri, k., "*Hydrodynamic optimization of propeller using Gradient based approach*". 14th Conference on Fluid Dynamics, Birjand, Iran, May 1-3, 2012

[22] Kerwin, J. E., "Hydorofoils and Propellers". MIT course, 13.04 lecture notes. http://ocw.mit.edu/courses/mechanical-engineering/2-23-hydrofoils-and-propeller-spring-2007

[23] Kimball, R. W., and Epps, B., "OpenProp v2.4 propeller/turbine design code", http://openprop.mit.edu, 2010

[24] S. S. Rao, Engineering Optimization, Theory and Practice. J, Wiely & Sons Inc, Fourth Edition, School Of Mechanical Engineering, University of Purdue, 2009

[25] G. N. Vanderplaats, Numerical Optimization Techniques for Engineering Design, 1984

[26] Brizzolara, S, Villa, D, and Gaggero, S. "A Systematic Comparison between RANS and Panel Method for Propeller Analysis". Proc. of 8th International Conference on Hydrodynamics, Nantes, France, 2008

[27] Mitchel, M., "An introduction to Genetic Algorithms". MIT Press, Cambridge, MA, 1998

[28] Coley, D. A., "An Introduction to Genetic Algorithms for Scientists and Engineers". World Scientific Publishing Co., Singapore, 1999

[29] Karim, M. M., Suzuki, K., and Kai, H., "Optimal Design of Hydrofoil and Marine Propeller Using Micro-Genetic Algorithm". Journal of Naval Architecture and Marine Engineering, December, 2004

[30] Aykut, S., Kentli, A., Gulmez, S., and Yazicioglu, S., "Robust Multiobjective Optimization of Cutting Parameters in Face Milling". Journal of Applied Science, Acta Polytechnica Hungarica, Vol. 9, No. 4, 2012

Diagnostic Measurement for the Effective Performance of Motor Vehicles

István Lakatos

Széchenyi István University
Department of Road and Rail Vehicle
Egyetem tér 1, H-9026 Győr, Hungary
lakatos@sze.hu

Abstract: Diagnostics means instrumental measurement without dismantling. So measuring the engines on rolling roads is not incorporated in this concept, as this measurement is accomplished on engines taken out from vehicles. During service, it is often necessary to measure the performance of the engines, or to judge the performance projection of some corrections. The new method discussed in this article provides a solution to this problem.

Keywords: drive train; wheel performance; effective performance; free acceleration

1 Introduction

The measurement of the effective performance of motor vehicles takes place with the aid of bench tests that suggest a dismantled engine. The results of this in case of engines are the usually available so called external or total load characteristics.

In the case of operating vehicles, the possibility to take such measurements is quite rare. During diagnostics and repair work it is necessary more and more frequently to measure the effective moment and performance of the engines.

Some rolling roads possess these kind of skills; however even their prices exceed the budget of the services.

Henceforth, I introduce the theoretical background of a new measurement method with accessible instruments [6].

2 Measuring Engine Performance on Free Rollers

2.1 The Theoretical Background of the New Measurement Method

The principle of the measurement method is to accelerate and decelerate the unloaded drive train of the studied vehicle on free rollers (there is no need for a rolling road). Since we want to measure external characteristics, the measurement must be performed under total load conditions [1], [2].

Let us write the energy equation for the system displayed in Figure 1; according to this equation the temporal change of the **introduced work** in the system (P_e – effective engine performance) equals to the temporal changes of the **kinetic energy** (E_k), the **potential energy** (E_p) and the **diverted heat** (Q):

$$P_e = \frac{dE_k}{dt} + \frac{dE_p}{dt} + \frac{dQ}{dt} \tag{1}$$

As the potential energy does not change during the measurement:

$$P_e = \frac{dEk}{dt} + \frac{dQ}{dt} \tag{2}$$

The kinetic energy change of the system is displayed in the acceleration of the wheel and the rollers, so this element equals to the wheel performance (P_k). Though the diverted heat equals to the running loss performance (P_v):

$$P_e = P_k + P_v \tag{3}$$

The basic dynamic equation of rotation can be prescribed both for acceleration and deceleration phases:

$$P = M \cdot \omega = (\theta_{red} \cdot \varepsilon) \cdot \omega = \theta_{red} \cdot \frac{d^2\varphi}{dt^2} \cdot \frac{d\varphi}{dt} \tag{4}$$

where:

- ω the angular velocity of the roller of the rolling road
- Θ_{red} moment of inertia of the drive train of the vehicle reduced to the shaft of the roller of the rolling road
- φ angular displacement of the roller
- ε angular acceleration of the rollers
- t time

The basic equation of the acceleration phase:

$$P_{\text{veszt}} = P''_{\text{v,f(M)}} + P_{\text{v,f}(\omega)} + P_{\text{vpad,f(vg)}} \tag{5}$$

$$P_{\text{kerék}} = \omega_g \cdot \varepsilon_g^+ \cdot [\theta_{\text{mot,red}} + \theta_{\text{járműáred}} + \theta_{\text{pad}}] \tag{6}$$

where:

- f(M) the f(M) index elements are the losses depending on the tractive force
- f(ω) the f(ω) index elements are the losses depending on the velocity
- $_g$-index roller
- $\Theta_{\text{mot, red}}$ value of the moment of inertia of the engine reduced to the shaft of the roller
- $\Theta_{\text{jármű, red}}$ value of the moment of inertia of the vehicle reduced to the shaft of the roller
- $\Theta_{\text{pad, red}}$ value of the moment of inertia of the rolling road reduced to the shaft of the roller
- $^+$-index acceleration
- $^-$-index deceleration

The basic equation of the deceleration phase:

$$P_{\text{veszt}} = P'_{\text{v,f(M)}} + P_{\text{v,f}(\omega)} + P_{\text{vpad,f(vg)}} \tag{7}$$

$$P_{\text{fékező}} = \omega_g \cdot \varepsilon_g^- \cdot [\theta_{\text{járműáred}} + \theta_{\text{pad}}] \tag{8}$$

In the aforementioned equations the alternations of $P_{v,f(M)}$ value marked with one or two commas refer to a different loss proportion in the acceleration and the deceleration phases, as in the latter phase the engine is separated from the system.

The phases of the measurement are the following:

1 **ACCELERATION PHASE:** The drive train of the vehicle on the bench and the rollers of the rolling road are accelerated, in the studied gear accelerated up to the rated engine speed with full load (on full blast).

2 **DECELERATION PHASE:** By releasing the clutch, leaving the gear at the given position, we let the car decelerate until it stops.

During the measurement, as there is no external load, the engine has to accelerate the moments of inertia indicated in Figure 1. During the deceleration we separated the moment of inertia of the engine, so with this exemption the rest of the moments of inertia decelerate the system.

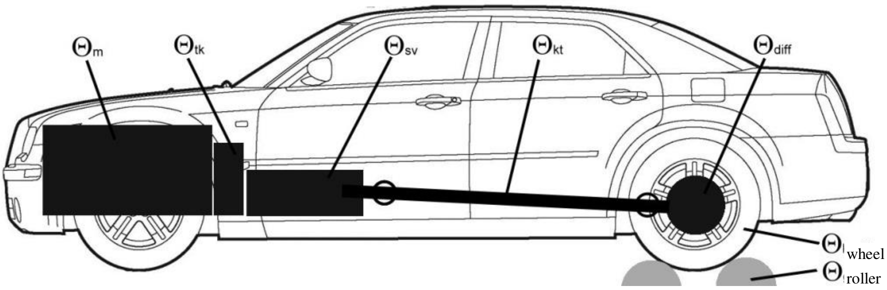


Figure 1
Vehicle drive train on free rollers

During the measurement only one transmitter is needed; we need to measure the revolution number (n) of the shafts of the rollers (roller radius: r_g). From this we can formulate the following data:

- angular velocity: $\omega = \frac{d\varphi}{dt} = 2 \cdot \pi \cdot n,$
- angular acceleration: $\varepsilon = \frac{d\omega}{dt}$
- vehicle speed (the circumferential velocity of the wheel, and the roller) $v = r_g \cdot \omega$

According to the aforementioned measurements the diagram displayed in Figure 2 can be recorded.

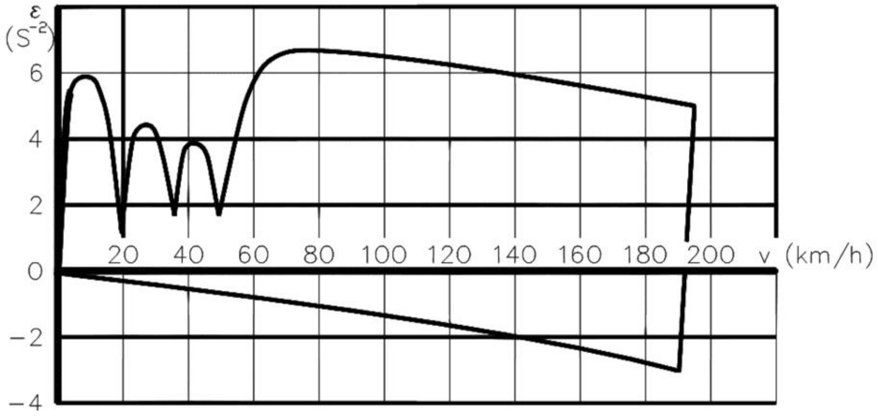


Figure 2
Characteristic curve recorded during the measurement

The following fundamental mechanic equations should be taken into further consideration:

$$M = \theta \cdot \varepsilon \tag{9}$$

and

$$P = M \cdot \omega = \theta \cdot (\varepsilon \cdot \omega) \tag{10}$$

In accordance with the aforementioned equations we mean by Θ the reduced moment of inertia on the shaft of the roller (Θ_{red}), as we also measure the revolution number on the shaft of the roller.

According to the aforementioned ideas we can state:

$$M(v) \sim \varepsilon(v) \tag{12}$$

$$P(v) \sim (\varepsilon \cdot \omega)(\mathrm{v}) \tag{13}$$

Namely, the ε function is in accordance with the full load torque curve as regards the features of vehicle speed (or even revolution number), while (ω.ε) function is parallel to the performance curve.

2.2 Practical Considerations of the Measurement

The performance rated functions can be seen in Figure 3. It must be interpreted before moving on:

1 **ACCELERATION PHASE**: the roller is accelerated by the wheel of the vehicle, thus the function taken with the measurement (ω.ε) is proportional with the wheel performance.

2 **DECELERATION PHASE**: The roller and the drive train are decelerated by the loss of the units after the engine, namely the loss performance.

Namely:

$$P_{wheel} = P_{eff} - P_{loss} \tag{14}$$

In reverse:

$$P_{eff} = P_{wheel} + P_{loss} \tag{15}$$

Where:

- P_{eff} – the effective performance of the engine
- P_{loss} – the loss of the drive train
- P_{wheel} – the performance deduced on the wheel

The practical realization of the latter function can be seen in Figure 3, where the summary of the proportionate functions about the performance of the wheels and the loss of the drive train are indicated. The resultant function is a proportionate curve with the (effective) performance of the engine.

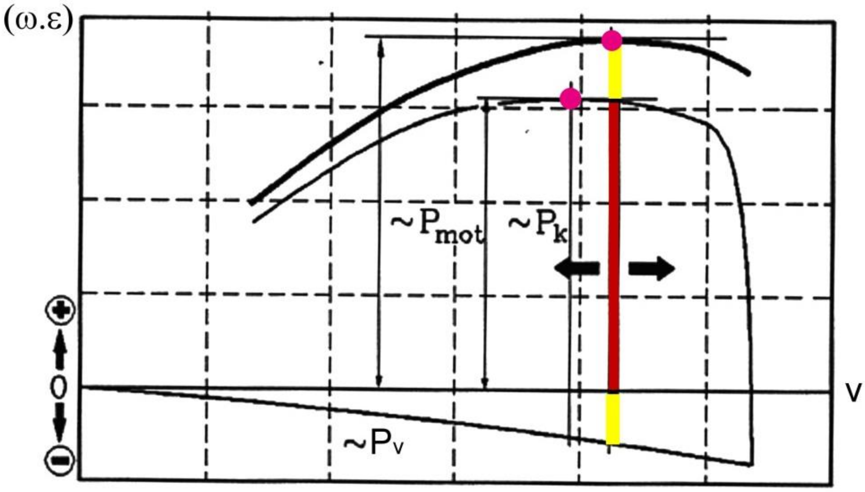


Figure 3
Performance proportionate characteristic curve recorded during the measurement

3 Evaluation of the Measurement Results

The main condition for evaluating the results of the measurement is to know the moment of inertia indicated in the Figure 1. Although these values are not available, their shortage can be eliminated with an adequate measurement procedure.

3.1 Validating the External Characteristic Curve of the Engine by Roller Bench Test

If we possess not only rollers but our measuring rollers are the rollers of a vehicle bench pad, we have to find the velocity value (v^*) that belongs to the local maximum of the acceleration phase indicated in Figure 3 [3], [4].

As a next step we choose the load characteristics of constant velocity on the roller bench, we set the v^* velocity, and then we measure the wheel performance with full load (Figure 4).

With these steps the values appear in kW even on the vertical axis of Figure 3, i.e. the effective performance curve of the engine is available.

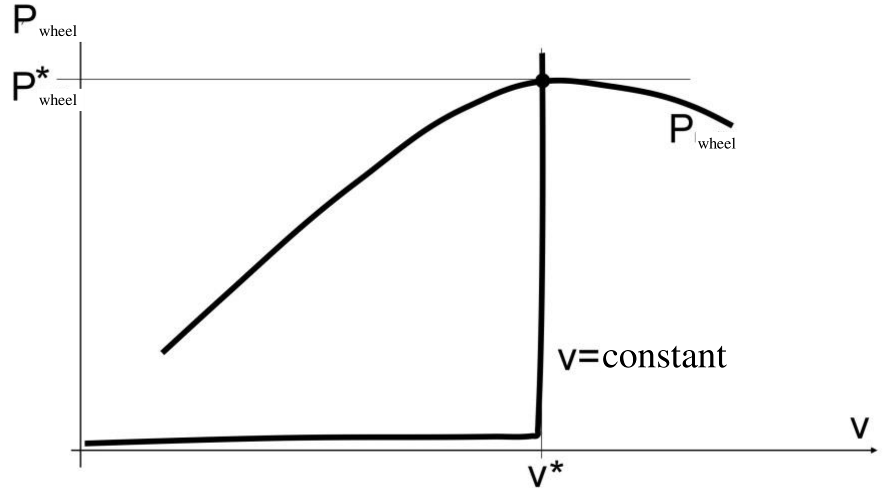


Figure 3
Performance proportionate characteristic curve recorded during the measurement

The flaw in this method is that the scales of the diagram phases of the acceleration and deceleration are not equal, as in one case we should take into consideration the (unknown) moment of inertia of the engine, while in the other case it is not needed. This fact can be considered only with a correction factor defined empirically (generally 10%).

3.2 Validating the External Characteristic Curve of the Engine without Roller Bench Test (New Measurement Method)

The main point of the measurement method which I elaborated is the following:

If we do not possess a rolling road, just a roller bed, we need to conduct two free acceleration measurements.

1st Measurement

Actually, this is equal with the previously mentioned measurement method. Equations:

$$P_{k,1} = (\theta_{engine,red} + \theta_{drive\ train,red}) \cdot \varepsilon_1 \cdot \omega_1 \tag{16}$$

$$P_{v,1} = \theta_{drive\ train,red} \cdot \varepsilon_1 \cdot \omega_1 \tag{17}$$

2nd Measurement

In this case we tie an additional flywheel to the shaft of the rollers, which enhances the total reduced moment of inertia calculated for the shaft of the roller.

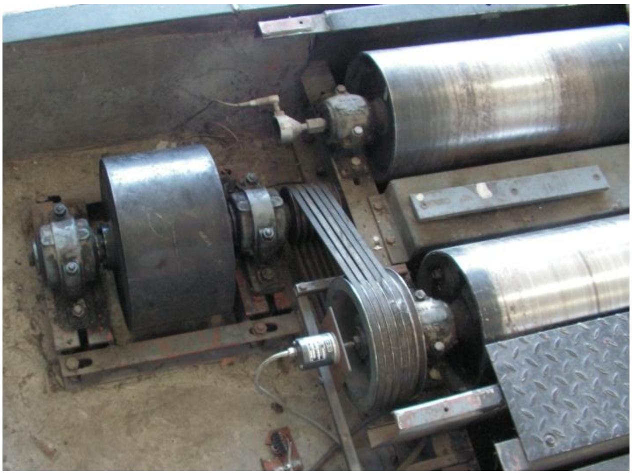


Figure 4
Roller bench with the rotating mass

$$P_{k,2} = (\theta_{engine,red} + \theta_{drive\ train,red} + \theta_{flywheel,red}) \cdot \varepsilon_2 \cdot \omega_2 \quad (18)$$

$$P_{v,2} = (\theta_{drive\ train,red} + \theta_{flywheel,red}) \cdot \varepsilon_2 \cdot \omega_2 \quad (19)$$

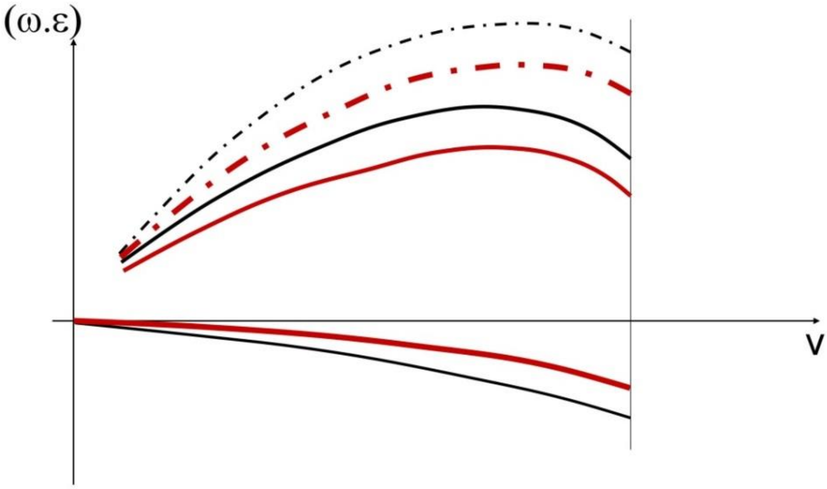


Figure 5
Diagrams of two subsequent measures

Although in both of these cases the inertia and naturally the acceleration and deceleration of the system are diverse, by calculating the performance the performances through the wheel and the performance loss must correspond with each other, since the engine that accelerates the system and the drive train that has losses are unaltered.

Consequently, in the case of the 1st and 2nd measurements, the equations of the wheel performance and the drive train loss can be equated in couples:

$$\begin{aligned}(\theta_{engine,red} + \theta_{drive\ train,red}) \cdot \varepsilon_1 \cdot \omega_1 \\ = (\theta_{drive\ train,red} + \theta_{flywheel,red}) \cdot \varepsilon_2 \cdot \omega_2\end{aligned} \qquad (20)$$

(20)

$$\theta_{drive\ train,red} \cdot \varepsilon_1 \cdot \omega_1 = (\theta_{drive\ train,red} + \theta_{flywheel,red}) \cdot \varepsilon_2 \cdot \omega_2 \qquad (21)$$

Thus the following is derived from the 2nd equation:

$$\theta_{drive\ train,red} = \frac{(\varepsilon_1 \cdot \omega_1 - \varepsilon_2 \cdot \omega_2)}{\theta_{flywheel,red} \cdot \varepsilon_2 \cdot \omega_2} \qquad (22)$$

In the above equation the numerator is measured, and the denominator is a datum known from the construction. On the basis of this, with the help of the 1st equation, even the moment of inertia of the engine can be determined.

All of this process is naturally controlled and calculated with the aid of adequately elaborated measuring software. Consequently, the end result will be an external characteristic curve of torque and performance defined by exact values.

Summary

There are several advantages to elaborating a new measuring system [5]:

1 There is no need for a rolling road, and therefore simpler and cheaper measuring devices can be developed.

2 The measurement is more precise, since it defines everything on the basis of measuring excluding empirical correction factor.

The described new method provides an available opportunity for professional services, as it is able to define the diagnostic performance of the engine with required exactness.

It is a very important achievement that with the help of diagnostic tools (without the removal of the engine) we can get an accurate result of the effective performance of the engine, as it significantly differs from the performance result of the wheels that can be measured on the rolling road (Figure 6).

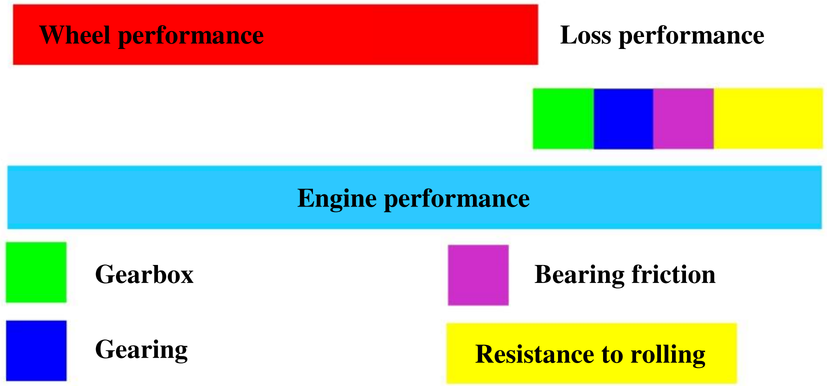


Figure 6
Wheel performance– drive train losses – effective performance of the engine

The value of these losses (Figures 7 and 8) cannot be defined with measuring or can only be defined with difficulty.

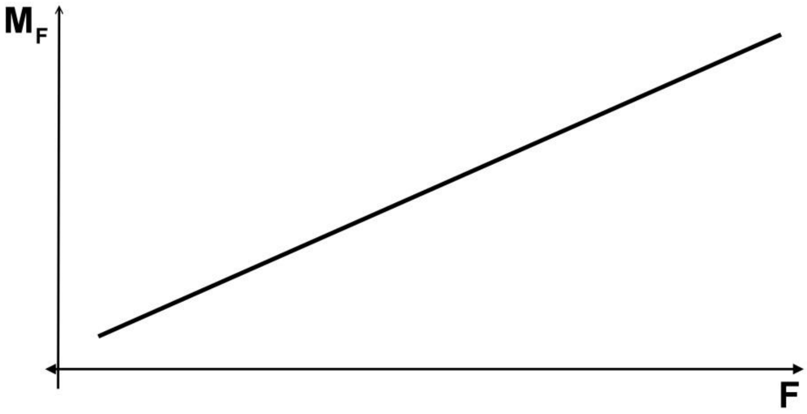


Figure 7
Loss depending on tractive force (Gear wheel friction, slip)

Ultimately the most significant result of the new procedure is that it enables us to measure the effective performance of the engine, for which we usually possess reference data, with adequate accuracy in a testing bench.

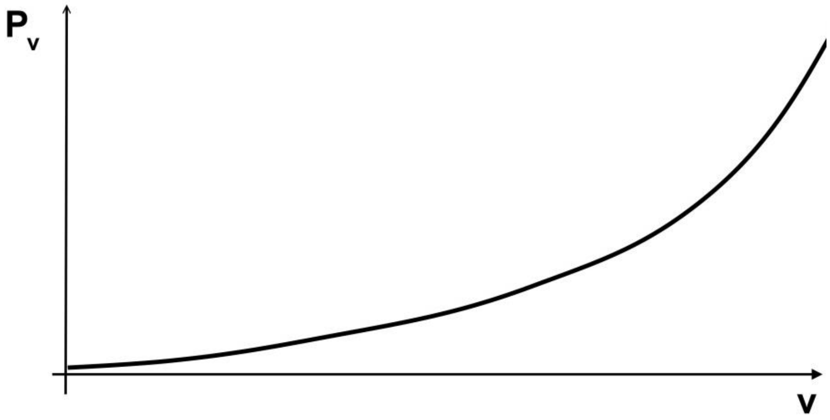


Figure 8
Loss depending on velocity (oil mixing, ventilation, tyre kneading work)

References

[1] Dr. Lakatos István: Analyse der Zusammenhängen zwischen indizierten Motorkennwertwen und Rollprüfstanduntersuchungsergebnissen, JÁRMŰVEK 49:(6) pp. 31-34. (2002)

[2] Dr. Lakatos I.: Comparative Measures on Rolling Road Dynamometers, XXV microCAD International Scientific Conference, 31 March – 1 April 2011, Miskolc, Hungary, pp. 57-64

[3] Dr. Lakatos I.: Gasoline Engine Diagnostic on Chassis Dynamometers, XIX microCAD International Scientific Conference, 18-20 March, 2010, Miskolc, Hungary, pp. 27-32

[4] Dr. Lakatos I.: Instacioner Engine Performance Measure on Rolling Road Dynamometers, XIX microCAD International Scientific Conference, 18-20 March, 2010, Miskolc, Hungary, pp. 33-38

[5] Dr. Lakatos I.: Measuring Engine Performance with Diagnostic Tools, Innovation and Sustainable Surface Transportation Conference, 3-5 September 2010, Budapest, Conference material on CD

[6] Dr. Lakatos I.: Motorteljesítmény mérés diagnosztikai eszközökkel In: Péter Tamás (ed.) Innováció és fenntartható felszíni közlekedés konferencia: IFFK 2010, Budapest, Hungary, Sept. 2-4, 2010, Budapest: BMF; Magyar Mérnökakadémia, 2010 (ISBN 978-963-88875-1-1) CD proceedings