INDEX 330930 ISSN 0860-8229

# M&MS

2017 1

# METROLOGY AND MEASUREMENT SYSTEMS

QUARTERLY, VOLUME 24

PAN POLSKA AKADEMIA NAUK

POLISH ACADEMY OF SCIENCES
COMMITTEE ON METROLOGY AND SCIENTIFIC INSTRUMENTATION

WARSAW 2017

**METROLOGY AND MEASUREMENT SYSTEMS**
**Quarterly of Polish Academy of Sciences**



Journal is indexed by Journal Citation Reports/Science. Impact Factor: 1.140 (5-Year Impact Factor 1.092).
More information about aims and scope of the journal – inner side of the back cover.
Instructions for Authors – last pages of the issue.
Edition was financially supported by the Polish Academy of Science and Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics.

Ark. wyd. 17,65 Ark. druk. 14,12
Papier offsetowy kl. III 80g 70 x 100 cm
Print run 120 copies

Druk: Wrocławska Drukarnia Naukowa PAN Sp. z o.o.
im. Stanisława Kulczyńskiego
53-505 Wrocław, ul. Lelewela 4

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 3–18.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# AUTOMATIC EXTRACTION OF THE PELVICALYCEAL SYSTEM FOR PREOPERATIVE PLANNING OF MINIMALLY INVASIVE PROCEDURES

**Katarzyna Heryan[1)], Andrzej Skalski[1)], Jacek Jakubowski[2)], Tomasz Drewniak[3)], Janusz Gajda[1)]**

*1) AGH University of Science and Technology, Department of Measurement and Electronics, Al. Mickiewicza 30, 30-059 Cracaw, Poland (✉ heryan@agh.edu.pl, +48 12 617 3596, skalski@agh.edu.pl, jgajda@agh.edu.pl)*
*2) Rydygier Memorial Hospital, Department of Urology, Os. Zlotej Jesieni 1, 31-826 Cracaw, Poland (jacekjakubowski83@gmail.com)*
*3) Specialized Municipal Hospital G. Narutowicz, Department of Urology, Prądnicka 35-37, 31-202 Cracaw, Poland (tomdrew@vp.pl)*

**Abstract**

Minimally invasive procedures for the kidney tumour removal require a 3D visualization of topological relations between kidney, cancer, the pelvicalyceal system and the renal vascular tree. In this paper, a novel methodology of the pelvicalyceal system segmentation is presented. It consists of four following steps: ROI designation, automatic threshold calculation for binarization (approximation of the histogram image data with three exponential functions), automatic extraction of the pelvicalyceal system parts and segmentation by the Locally Adaptive Region Growing algorithm. The proposed method was applied successfully on the Computed Tomography database consisting of 48 kidneys both healthy and cancer affected. The quantitative evaluation (comparison to manual segmentation) and visual assessment proved its effectiveness. The Dice Coefficient of Similarity is equal to $0.871 \pm 0.060$ and the average Hausdorff distance $0.46 \pm 0.36$ mm. Additionally, to provide a reliable assessment of the proposed method, it was compared with three other methods. The proposed method is robust regardless of the image acquisition mode, spatial resolution and range of image values. The same framework may be applied to further medical applications beyond preoperative planning for partial nephrectomy enabling to visually assess and to measure the pelvicalyceal system by medical doctors.

Keywords: the pelvicalyceal system segmentation, kidney segmentation, kidney compartments, Computed Tomography (CT), kidney cancer.



## 1. Introduction

In 2013 in Poland, there were 5178 new cases of kidney cancer reported. The incidence rates for males and females reached respectively 16.8 and 10.3 per 100.000 [1]. Widely used imaging techniques, especially ultrasonography, cause that nowadays a majority of new renal tumours are detected at an early stage. This provides an opportunity for removing only the cancer lesion and preserving as much of the healthy kidney parenchyma as possible (*nephron sparing surgery* – NSS, *partial nephrectomy* – PN). In recent years, this approach has become a gold standard in treatment of small renal masses. The NSS offers better functional results comparing with the *radical nephrectomy* (RN), with an acceptable profile of peri- and postoperative complications and a similar long-term oncological outcome.

Although the decision regarding treatment feasibility must include all clinical aspects (the patient's age, general conditions, comorbidities, *etc*.), the preoperative planning based on thorough image analysis is essential. It is particularly important to correctly assess the tumour extent and its boundaries, to distinguish all surrounding structures within and beyond the kidney and to examine their mutual topological relations in order to identify possible conflicts in the operated area. This assessment may be based on conventional 2D CT scans, but it can be also supported by 3D reconstructions together with other post-processing techniques, such as

Article history: received on Jun. 07, 2016; accepted on Jan. 15, 2017; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0014.

various types of rendering and segmentation. It was suggested that, in preoperative planning or intraoperative navigation, automated segmentation can outperform rendering, as it can produce semi-transparent dynamic models grouping information obtained from different phases, in comparison with limited to a single phase opaque pictures offered by rendering alone [2]. The approach based on image analysis meets the latest trends in surgical oncology. One of its key stages is a 3D segmentation of anatomical structures (kidney, the *pelvicalyceal system* (PCS), vascular tree). In this paper we focus on the PCS segmentation.

According to the European Association of Urology Guidelines on *Renal Cell Carcinoma* (RCC) [3] both CT and MRI are used to characterise the renal mass. The diagnostic decision on RCC is made on the basis of a change of 15 HU or more on the depicted renal mass on CT scans before and after administration of a contrast, preferably in the nephron-graphic phase to maximise differential diagnosis and detection. If the results of CT scan are ambiguous, then MRI may be considered. Additionally, the triple-phase abdominal contrast-enhanced CT provides useful and detailed information on the renal vascular tree and PCS. Although the MRI tissue resolution is considered as higher than CT's, the possibility of acquiring all required information for surgery planning during one examination led to the fact that a triple-phase abdominal CT scan became a standard diagnostic procedure preceding the renal cancer treatment. Therefore, our research material consisted of triple-phase abdominal contrast-enhanced CT scans.

The major challenge of the PCS segmentation on a CT scan is the lack of synchronization between the contrast agent flow and the image acquisition time, which results in visualization of undesired structures and poor recognition of the target ones. Also, the intensity values within the same structure may vary (PCS inhomogeneity). These inconsistencies impact data processing and require a robust segmentation method that can automatically handle different acquisition protocols.

Since the presented issue is relatively new, there are only a few papers concerning the segmentation of kidney compartments. Most of the previously proposed solutions refer to MRI. The algorithms focus on the kidney segmentation without giving particular attention to the PCS. A comprehensive review of the algorithms proposed for kidney volume assessment on MRI provided in [4] presents different acquisition techniques and tailored segmentation methods. Will *et al.* [5] proposed a renal cortex, medulla and renal pelvis segmentation from MRI by simple thresholding. In the case of cortex, the classification is based on a threshold calculated as a row's mean inside the kidney mask minus a standard deviation. With regard to the image value variation within the same structure (explained in Section 2) this kind of solution cannot be applied to the PCS segmentation. Yang *et al*. [6] after kidney segmentation from DCE-MRI data used Principal Component Analysis to describe the data. Based on this, the k-means classification was applied to label data into particular structures. In the last stage, noise and misclassification problems were reduced. They achieved – according to *Dice Coefficient of Similarity* (DICE) – the result of renal pelvis (a part of PCS, Fig. 1) segmentation equal to 0.69 for disordered kidneys and 0.95 for healthy kidneys. Li *et al*. [7] proposed renal cortex segmentation on the CT data using an optimal surface search method. The quality of multiple surfaces' searching method was improved by a graph construction scheme. The segmentation effectiveness was equal to $0.741 \pm 0.032$ in the true positive volume fraction (precision) and $0.0008 \pm 0.0013$ in the false positive volume fraction (1-specificity).

Furthermore, we can distinguish two major groups of segmentation algorithms for kidney or/and its compartments: *region growing* (RG) and *level set* (LS) methods. In the RG group three papers are worth mentioning. Pohle *et al*. [8] developed a 2D *adaptive region growing* (ARG) method consisting of two steps. In the first one, the region inhomogeneity is iteratively estimated starting from one seed pixel and after each region doubling the ad hoc coefficients are used to compensate for constant underestimation of the standard deviation. The estimated

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 3–18.

parameters (median and both upper and lower standard deviations) are then used in the second RG step. Important assumptions of this method cover the requirement of region compactness, the seed location not on the structure boundary and a sufficient number of pixels in the region. The second paper, written by Abiria *et al*. [9], concerns an RG algorithm (EdgeWave) constrained by a user-specified morphological erosion radius and edge strength. The seed set can be defined by thresholding the selected region, which is facilitated by the average value of a selected region that is displayed in a control window. In the third paper [10], due to the false assumption that there are only two types of renal perfusion signals in the object area, the RG is used to find an approximate solution to Chan-Vese (ChV)-based energy functional [11] to overcome this. The RG is driven by the correlation coefficient quantifying the similarity between the average temporal sequences of two adjacent regions. Among the LS methods developed for kidney segmentation, two can be distinguished [12, 13]. The first one applies a 3D kidney segmentation to DW-MRI images and is based on an adaptive shape, prior guided by the first- and second-order visual appearance features of DWI-MRI data. The kidney and the background models are provided by integrating these features into a joint Markov-Gibbs random field. Although this method is well designed for the kidney segmentation, it is inadequate for the PCS segmentation due to a variety of PCS shapes requiring a huge database to create a reliable model. In the second paper, a 4D LS framework combining both spatial and temporal information is proposed to segment the cortex, medulla and collecting system from a dynamic MRI. Again, since our material is included in a CT database, the temporal information cannot be provided. Although the LS algorithms have been proposed for the MR data and have taken into account their specificity, the common denominator of these two methods is the usage of the ChV term [11].

In this paper we propose an automatic PCS segmentation method on the delayed CT data phase. The presented methodology is based on our previous work – the *Locally Adaptive Region Growing* technique (LARG) [14] driven by an image value with the automatic initialization stage. In order to provide a reliable assesment of the proposed method, its comparison with ARG [8], EdgeWave [9] and ChV [11] methods was performed on the same CT database.

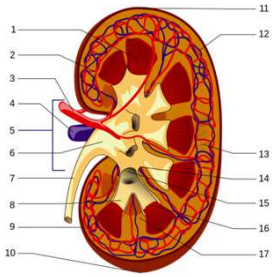
This paper is constructed as follows: Section 2 presents the CT database description together with associated challenges to be addressed; Section 3 highlights the medical importance of PCS segmentation; the proposed segmentation methodology is described in Section 4; the results and conclusion are discussed in Sections 5 and 6, respectively.

## 2. Material and associated challenges

The research material consisted of 24 abdominal CT scans that were acquired in standard medical procedures preceding the oncological surgery. The studies were performed in various medical centres and on different devices. Individual studies differed in the acquisition mode and using various protocols. This resulted in acquiring a set of data that vary in their image intensity values and spatial resolution (spacing: 0.609−0.885 mm, slice thickness 1−5 mm). The dataset for each patient consisted of three phases (the arterial, venous and delayed ones). In each phase the contrast passes different compartments of the kidney. In the delayed phase most of the contrast agent should be excreted in urine, depicting selectively the urinary tract (including the PCS). Therefore, this phase was chosen for the PCS segmentation. In detail, the PCS segmentation process involves segmentation of several structures that are indicated in Fig. 1. For the sake of simplicity, we refer to all of them as the PCS.

The major challenge is the lack of synchronization between the contrast agent flow and the image acquisition time. This results in visualization of undesired structures and poor recognition of the target ones. For instance, at the beginning of excretory (delayed) phase there is still a high concentration of contrast agent in collecting ducts making papillae barely

distinguishable from calyces. Also, the intensity values within the same structure may vary (from this point of view the PCS is not a homogenous structure).



1. Renal pyramid
2. Interlobular artery
3. Renal artery
4. Renal vein
5. Renal hilum
6. Renal pelvis
7. Ureter
8. Minor calyx
9. Renal capsule
10. Inferior renal capsule
11. Superior renal capsule
12. Interlobular vein
13. Nephron
14. Renal sinus
15. Major calyx
16. Renal papilla
17. Renal column

Fig. 1. The kidney compartments. The parts of the PCS are underlined; [F1].

To ensure proper validation of the segmentation results the manual segmentation was performed (48 outlines of PCSs), using the ITK-SNAP software [15], by an experienced urologist. Although the PCS is a continuous structure, in some of the manual outlines there are visible gaps, most commonly in the calyceal neck topography and in the transition between renal pelvis (Fig. 1.6) and ureter (Fig. 1.7). These structures are the narrowest sections of the upper urinary tract. The calyceal neck width normally ranges from 2 to 4mm in 17–48% of patients, depending on the group of calyces, and in the case of 3% of patients the width of the central calyceal is below 2mm [16]. Thus, due to an insufficient transversal resolution and/or thickness of the scanned slices (here, up to 5mm), these structures are incompletely depicted in a CT scan. The discontinuous representation of PCS is another challenge that affects both the accuracy of manual outlines and the effectiveness of segmentation. The situation is also worsened by two following reasons. The PCS is a complex, tree- like structure, which makes the process of manual segmentation of branches into 2D slices more difficult (sometimes a branch is only partially visible on one slice). However, it can be resolved by tracking the outlines simultaneously in other planes (coronal and sagittal), which is extremely time-consuming. The second issue is the slice visualization mode selected for the manual segmentation. Fig. 2 shows how various window setups (*window level* – WL and window width – WW) lead to different manual outlines. The way of performing the manual segmentation has a significant impact on its volume and therefore on the automatic segmentation evaluation. Moreover, the same window setups may give different results depending on an image acquisition protocol (different intensity range and image resolution). Standardization of this process is still a matter of debate.

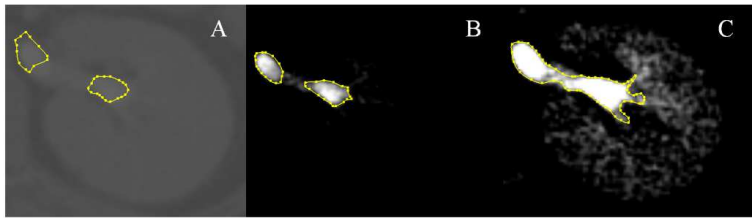



Fig. 2. An example of visualizing CT slices with different windows' parameters; (A) WW = 4137, WL = 1023; (B) WW = 216, WL = 238; (C) WW = 101, WL = 143. The influence of visualization parameters on manual contouring results (marked with a solid line) can be seen.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 3–18.

## 3. Medical importance of PCS segmentation

Besides preoperative PN planning, the anatomy of PCS is one of the crucial elements assessed, basing on imaging studies, during preoperative planning in many other kidney surgeries. Therefore, in order to facilitate its assessment and measurement, a proper segmentation is necessary.

In the surgical treatment of kidney cancer, a crucial parameter is the distance between tumour and the PCS as whole. It corresponds to the risk of opening the PCS and postoperative complications following unsuccessful reconstruction (urinary fistula). In treatment of kidney stones, the information about both shape and spatial configuration of individual components of the PCS (minor and major calyx and renal pelvis, Fig. 1: 8, 15, 6, respectively) are more important, especially when deposits are located in the lower group of calyces. This specific location hinders the possibility of reaching them during flexible ureterorenoscopy (RIRS) and impairs evacuation of remnants after *extracorporeal lithotripsy* (ESWL).

In the late 80s Sampaio [17] published an analysis of a large series of corrosive endo-casts of cadaveric kidneys' PCSs, determining particular calyceal patterns (anatomical variants) and their distribution. His anatomical classification remains the most widely used system in the preoperative assessment in endo-urology and ESWL. A few years later, together with Aragao [18], he published a paper indicating a correlation between the parameters describing the anatomy of the lower group of calyces (calyceal pattern, pelvicalyceal angle, length and width of the lower calyceal neck) and the effectiveness of ESWL. Since then, a large number of studies were carried out on the role of the PCS anatomy and the efficacy of ESWL/RIRS.

The majority of authors [19–22] use original parameters developed by Sampaio; some of them use their modifications, *e.g.* Elbahnasy [23], whereas others developed new ones, *e.g.* Fong [24]. Among all three aforementioned parameters, the pelvicalyceal angle remains the most important predictor of ESWL efficacy. The European Association of Urology guidelines [25] discourage using ESWL if the lower calyx (containing deposit) is either long (> 3 cm) or narrow (< 5mm), and the pelvicalyceal angle is steep (below 90 degrees). Although these parameters reflect spatial 3D configurations, they are routinely determined, basing on plain urography.

It has been reported that using the CT urography with a 3D reconstruction may increase the accuracy of PCS morphometry assessment, because it enables to measure its parameters in the actual PCS plane, not in a projection on the coronal plane, which is the case in plain urography [26–27]. In *percutaneous nephron-lithotomy* (PCNL), specification of an adequate access to the collecting system represents the most important success factor. It can be facilitated by using a 3D model of a particular PCS.

To meet the requirements of the described procedures and provide accurate measurements, a proper PCS segmentation on 3D CT images is essential.

## 4. PCS segmentation

The problems and challenges described in Section 2 influence the development of a segmentation method. Due to these problems we have proposed a method aiming at overcoming them (Fig. 3).

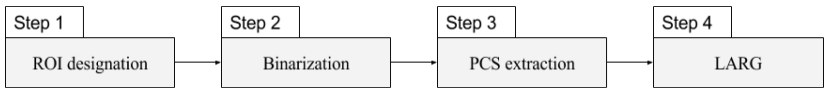



Fig. 3. An outline of the proposed method. A description is provided in the text.

At the beginning, the *region of interest* (ROI) (in this case, the kidney area) is determined (1st step). In the 2nd step binarization is performed in order to acquire the starting seed points for further segmentation. In the 3rd step the PCS parts and ureter are extracted from the binarized ROI image. The final segmentation is performed in the 4th step by the LARG algorithm [14], proposed by the authors.

### *4.1. Designation of region of interest (ROI), step 1*

The 1st step involves manual designation of ROI by indicating the kidney area (maximum span in each anatomical plane) with an adequate margin. It is very important not to crop the kidney as the image is also used for the kidney segmentation, therefore the margins around the kidney reach up to a few pixels (2–3). Additionally, cropping the whole image to ROI enables to significantly reduce the calculation time.

### *4.2. Binarization of ROI image (BW), step 2*

Due to the lack of study protocol standardization (various medical centres, devices, acquisition modes and, therefore, different image spatial resolution and value ranges) it is impossible to explicitly identify universal parameters, in particular the threshold value used for LARG initialization. As a consequence, it becomes necessary to develop an automatic method for accurate threshold determination for each individual case. A characteristic distribution of image values (Fig. 4) enabled to establish the following solution. It is inspired by the method proposed in [28], where a histogram of ROI image values is approximated by the sum of Gaussian functions. The main difference between this and our approaches lies in the way of threshold selection. In [28] the optimal threshold is set at the intersection of particular normal distributions. In the proposed method, the initial segmentation threshold is defined based on the normalized histogram of ROI image values approximated by the sum of three exponential functions:

$$F = \sum_{i=1}^{3} H_i \exp\left(-\frac{(x-a_i)^2}{2b_i^2}\right). \tag{1}$$

The function $F$ (1) is designated by a nonlinear approximation using the Trust-Region-Reflective Least Squares Algorithm [29]. Such an approximation was performed for each of 48 PCS images. In each case, the information extracted from $F$ was used to designate thresholds. The determined image intensity values of $a_i$ were sorted in the ascending order : $a_1, a_2, a_3$, together with corresponding values of $b_i$, so that $a_3 = \max(a_i)$. In each case, a threshold was set to $A \cdot a_3 + B \cdot b_3$. In order to calculate parameters $A$ and $B$, the dataset was divided randomly into two halves constituting the training set and the test set. Then, an optimization procedure, using the Nelder-Mead simplex direct search [30] aimed at minimizing the negation of the sum of DICEs calculated for the training set, was performed. The designated parameter values were as follows: $A = 1.006$, $B = 2.326$. Next, during the remaining steps of the proposed method (3rd and 4th step) only the test set was used. Additionally, to examine the dispersion of $A$ and $B$ parameters' values, a 10-fold partition of the training set was performed and $A$ and $B$ values were calculated independently for each fold. The experiment revealed that the coefficients did not differ significantly for the subsequent folds, as their values were equal to $1.005 \pm 0.002$ for $A$ and $2.353 \pm 0.142$ for $B$. Examples of approximations of the ROI image data by exponential functions are shown in Fig. 4. The calculated threshold value is indicated on the presented plots. Also, the variety of image intensity values depending on the patient CT data can be compared based on these two figures.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 3–18.

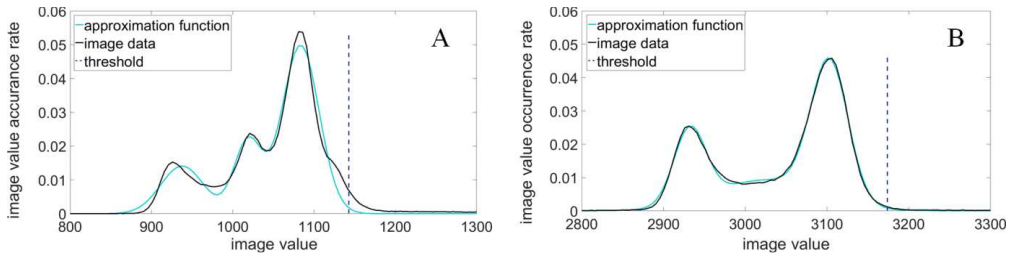



Fig. 4. Histograms of ROI image data normalized with an approximation function, two cases with different image values; $a_3 = 1083.469$, $b_3 = 22.778$ and *threshold* = 1142.95 (A); $a_3 = 3102.805$, $b_3 = 22.792$ and *threshold* = 3174.44 (B).

Since the desired structure – the PCS – is located inside the kidney, the goal of binarization is to separate it from the kidney parenchyma. The threshold should be high enough to achieve this separation even at the cost of discontinuity and loss of some PCS parts. Therefore, this step increases the fragmentation of PCS. For 2 examples from the test set the threshold had to be manually increased to the kidney parenchyma due to over-segmentation.

### *4.3. Extraction of PCS parts from binarized ROI image BW, step 3*

Within an ROI, after the preliminary segmentation, there often occur undesirable elements such as parts of spine or ribs. These elements have image intensity values at a comparable level as the structures with a contrast agent, which limits the possibility of indicating desirable structures based only on their values. In addition, discontinuities of the PCS after the preliminary segmentation preclude designation of the structure of interest by indicating one point located within the PCS. Fig. 5 explains the aforementioned issues (the outcome of initial segmentation – 2nd step).

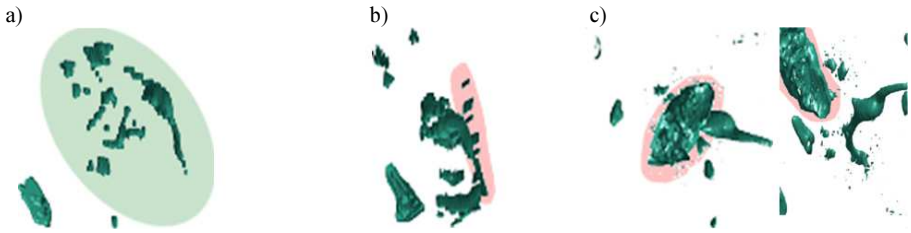



Fig. 5. Examples of initial segmentation results (2nd step) with indication of occurring problems. Discontinuities of the PCS (light green area) (a); close proximity between the PCS and spine (red area) (b); close proximity between the PCS and rib (red area), with transversal view (c).

Taking into account both the unfavourable mutual spatial location of desired and undesired structures and discontinuities of the PCS, we propose a *Hybrid Level Set* (HLS) method [31] to overcome it. The HLS is a modified classical ChV algorithm. In [32], among others, an additional prior term of ellipsoidal shape is introduced to keep the segmentation result within a desired range and to prevent it from leakage to the surrounding structures. This method was used for kidney segmentation and therefore enabled to identify the structures located inside it. Individual structures were recognized as the PSC when at least their portions were located inside the kidney. Visualizations of kidney segmentation results are shown in Fig. 6, the PCS parts and undesired structures are marked. Examples of the PCS extraction results – E, together with indication of undesirable structures after binarization (BW, 2nd step), are presented in Fig. 7.

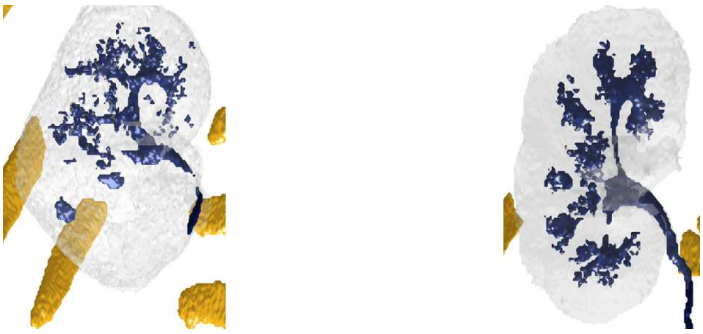


Fig. 6. Examples of results after step 3: kidney segmentation by the HLS (light grey color) with indicated structures located outside the kidney (orange) and extracted PCS (dark blue).

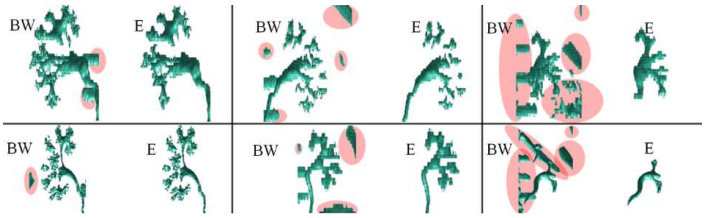



Fig. 7. Examples of the PCS extraction results. In each case the first picture indicate results after the binarization (BW, 2nd step, undesired structures are highlighted bya red area), and the second one – the results of extraction process (E, 3rd step).

### *4.4. Segmentation by Locally Adaptive Region Growing Algorithm, step 4*

The LARG algorithm [14] was developed and implemented as a remedy to the problems associated with the lack of homogeneity of voxel intensity values due to an improper contrast propagation. The method described in [14] was adjusted to meet the requirements of PCS segmentation. In comparison with the method proposed in [14], dedicated to vessel segmentation, the regularization was introduced and the region growing conditions were adopted to the PCS segmentation. In the proposed solution the conditions for voxel selection are specified individually for each candidate basing on analysis of its surroundings. The criteria for addition of a next voxel to the previously chosen ones are based on such parameters as the maximum permissible difference between the candidate voxel value and the source of growth *tDiff* (4) and the minimum required intensity value for a voxel candidate *tMin* (5). For data segmentation enhanced by a contrast agent only the lower threshold is meaningful. An outline of LARG algorithm is presented in Fig. 8.

At the beginning the default values are set such as:

– the maximum permissible difference between the candidate voxel value and the source of growth:

$$defaultDiff = 2 \cdot b_3; \tag{2}$$

– the minimum required value:

$$defaultMin = threshold - 0.5 \cdot b_3. \tag{3}$$

The parameters *threshold* and $b_3$ are derived from the 2nd step of the proposed method. This ensures the stability of LARG process providing individual parameter values adequate to a particular image. Since the default parameter values are used only if there are not enough

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 3–18.

already chosen voxels, a candidate should be examined basing on the restrictive conditions. Therefore, the *defaultMin* is set slightly below the preliminary segmentation threshold (2nd step) and *defaultDiff* – to $2 \cdot b_3$.

The initial seed set consists of voxels from the preliminary segmentation (3rd step, extraction). A set of seed values is sorted in the descending order to reduce the calculation time. Starting from the highest value (primary voxel) a region is growing in specified conditions. For each candidate the parameters *tDiff* and *tMin* are calculated basing only on previously added voxels. Starting from a radius equal to 1 (a cube (3,3,3)) it is checked whether there are enough previously classified voxels as objects to perform calculations. The size of the cube's neighbourhood is gradually increased until either the above condition is met or the maximum permissible size of the surroundings is achieved (max radius 5). The maximum radius is set to 5 to cover the standard diameter of the PCS structure (proper for the CT database transversal plane resolution). For each candidate the parameters are calculated as follows:

$$tDiff = \min\left(3 \cdot \mathrm{std}\left(\mathbf{tmpCT}\right) / radius, defaultDiff\right), \tag{4}$$

$$tMin = \max\left(\left(\mathrm{mean}\left(\mathbf{tmpCT}\right) - \mathrm{std}\left(\mathbf{tmpCT}\right) \cdot radius\right), defaultMin\right), \tag{5}$$

where **tmpCT** is a sub-image of a size defined by the radius and centre of the candidate's location.

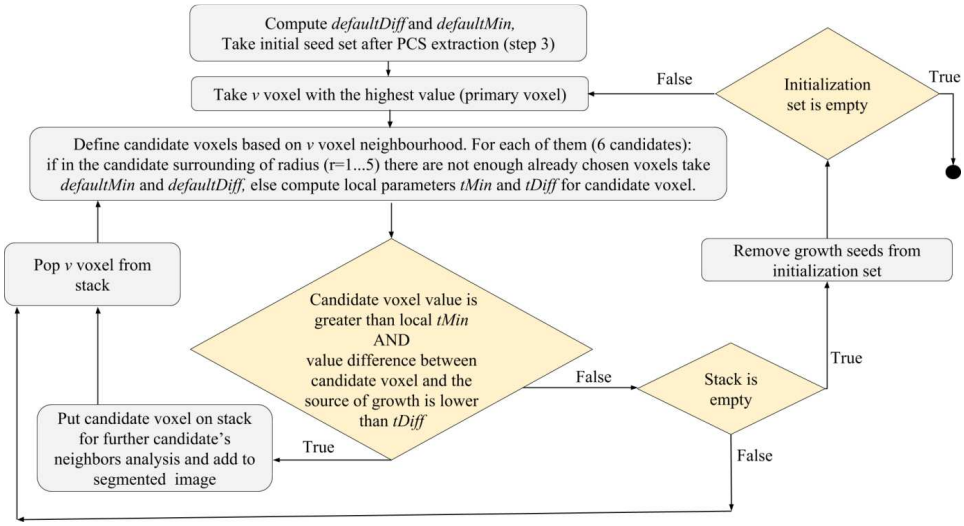



Fig. 8. The LARG – 4th step, an outline of algorithm.

If the number of previously added voxels is not required, the default values: *defaultDiff* (2) *defaultMin* (3) are used. If the parameters were calculated on the basis of a neighbourhood with a relatively large size, we put less confidence to their values (to a smaller extent they reflect the actual threshold for a candidate). Therefore, the regularization, that provides more stringent requirements for voxels located farther away from the candidate, is introduced. Also, the distant voxels may produce a greater variation of values and thus a relatively high *tDiff* (4). It has been observed during experiments that when disregarding the influence of neighbourhood size, the effectiveness of segmentation decreases. In some cases, this leads to overflow of the neighbouring structures. A candidate is added if its value is greater than *tMin* and the difference between the candidate and the source of grow is smaller than *tDiff* (4).

Starting from the primary voxel, voxels are put on the stack and subsequently removed after their growth ends. After emptying the whole stack, chosen voxels are removed from the initial seed set. The following step involves choosing a next primary voxel again. The operation is executed until the initial set of seeds is empty.

There are several advantages of defining region growing conditions locally, instead of using global parameters. First of all, it enables to add voxels whose intensity values descend towards the boundaries and along the vessel branch. Secondly, the voxels with lower intensity values, due to improper contrast propagation, may be also included. Furthermore, it prevents from leakage to unwanted structures. Finally, the continuity of the obtained structure is improved.

## 5. Results and discussion

Quantitative evaluation of the results from the test set (24 PCSs) was carried out on the basis of the following volumetric metrics derived from a confusion matrix (explanation in Table 1): accuracy, precision, sensitivity, specificity, over- and under-segmentation rate and DICE [33]. Apart from the volumetric metrics, also the distance measures: the maximum (*mHD*) [34] and the averaged Hausdorff distance (*aHD*) [35] were calculated. In general, the Hausdorff distance determines how far two subsets of a metric space are from each other. In particular, *mHD* denotes the maximum distance and *aHD* the average distance between points belonging to two subsets. The *aHD* distance is more suitable for image processing purposes as it decreases the impact of outliers [35]. Assuming that $\mathbf{P}_{mc}$ is a manual segmentation binary mask and $p_{mc}$ is a set of its points' coordinates, $\mathbf{P}_{seg}$ – the binary mask obtained by the algorithm and $p_{seg}$ – a set of its points' coordinates, the aforementioned measures (*mHD* and *aHD*) can be formulated as:

$$mHD = \max\left(h_1\left(\mathbf{P}_{mc}, \mathbf{P}_{seg}\right), h_1\left(\mathbf{P}_{seg}, \mathbf{P}_{mc}\right)\right), \tag{6}$$

$$aHD = \frac{h_2\left(\mathbf{P}_{mc}, \mathbf{P}_{seg}\right) + h_2\left(\mathbf{P}_{seg}, \mathbf{P}_{mc}\right)}{2}, \tag{7}$$

where: $h_1\left(\mathbf{P}_{\mathrm{mc}}, \mathbf{P}_{\mathrm{seg}}\right) = \max_{p_{mc} \in \mathbf{P}_{mc}} \min_{p_{seg} \in \mathbf{P}_{seg}} \left\| p_{mc} - p_{seg} \right\|$, $h_2\left(\mathbf{P}_{\mathrm{mc}}, \mathbf{P}_{\mathrm{seg}}\right) = \operatorname{mean}_{p_{mc} \in \mathbf{P}_{mc}} \min_{p_{seg} \in \mathbf{P}_{seg}} \left\| p_{mc} - p_{seg} \right\|$ and $\| \ \|$ denotes the Euclidian distance.

In simple words, the Hausdorff distance specifies how far (meaning a spatial distance) the acquired result boundary is from the manual outline. In order to denote the Hausdorff distance, first corresponding points in two subsets (the resulting boundary points and the expected ground true) are identified. Next, the distances between them are calculated. *mHD* is the longest distance among them and *aHD* is the average one. In this paper, the scores (6), (7) are provided in mm, presenting the distance between the segmentation result and the manual outline with regard to the image resolution.

The proposed method was applied successfully in each of 24 PCSs belonging to the test set. Both visual and quantitative analyses were accomplished to evaluate its performance. A visual comparison between manual outlines and the segmentation results is shown in Fig. 9. The provided examples indicate a good quality of the segmentation together with a substantial improvement achieved when applying LARG after PCS extraction. The visual assessment is supported by DICE provided for each case. The results are presented in Table 2. After the 3rd step of the proposed method the mean DICE is above 84%. Further improvements are achieved when applying the 4th step – LARG (87%). The calculated average Hausdorff distance (0.46 ± 0.36 mm) is satisfactory from a medical point of view.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 3–18.

Table 1. A confusion matrix together with definitions of evaluation metrics: accuracy, precision, sensitivity, specificity, DICE, over-segmentation and under-segmentation rates.

| | | **Actual value** | | | |
|---|---|---|---|---|---|
| | | object | background | | |
| **Predicted value** | object | **TP** (true positive) | **FP** (false positive) | **Precision (Prec)** =TP/(TP+FP) | **Over-segmentaton (OS)** = FP/TP |
| | background | **FN** (false negative) | **TN** (true negative) | **Accuracy (Acc)** =(TP+TN)/ (TP+FP+FN+TN) | **Under-segmentation (US)** = FN/TP |
| | | **Sensitivity (Sens)** =TP/(TP+FN) | **Specificity (Spec)** =TN/(FP+TN) | **DICE** =2TP/ (2TP+FN+FP) | |

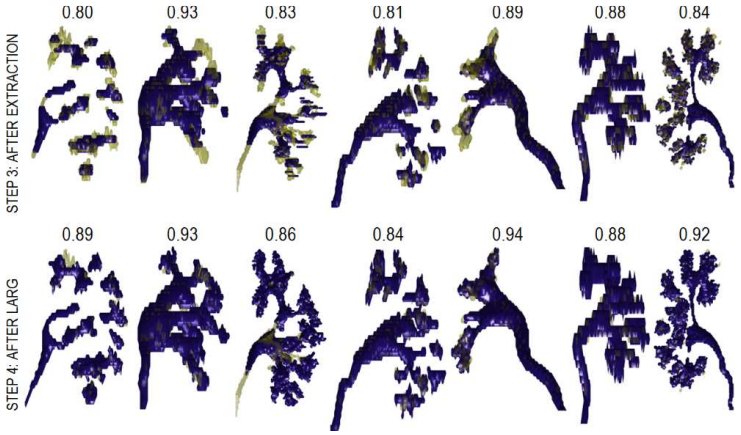



Fig. 9. An example of visualization of the obtained results in a 3D view. The rows indicate two subsequent steps of the proposed method, the columns represent 7 different cases.The DICE is presented for each case and each step. The dark blue color represents the output of the algorithm while the green one – the manual outline.

Table 2. A summary of the results indicating improvements after applying subsequent steps of the proposed method.

| | **Acc** | **Prec** | **Sens** | **Spec** | **OS** | **US** | **DICE** | ***mHD* [mm]** | ***aHD* [mm]** |
|---|---|---|---|---|---|---|---|---|---|
| **Step 3: After Extraction** | 0.997 ±0.004 | 0.945 ±0.063 | 0.778 ±0.114 | 0.999 ±0.001 | 0.070 ±0.079 | 0.315 ±0.219 | 0.845 ±0.068 | 9.96 ± 7.60 | 0.74 ± 0.49 |
| **Step 4: After LARG** | 0.997 ±0.004 | 0.872 ±0.107 | 0.883 ±0.074 | 0.998 ±0.003 | 0.166 ±0.165 | 0.141 ±0.101 | 0.871 ±0.060 | 7.73 ± 6.67 | 0.46 ± 0.36 |

The comparative assesment was also performed. The results of the proposed method were compared with those obtained with EdgeWave [9], ARG [8] and ChV [11] methods. For this purpose both ARG and ChV methods were implemented, while the EdgeWave tool was shared with its authors. The ARG was adapted for the 3D case. To assure the assumptions of this method (region compactness) each separatated part of the PCS was estimated and grown separately. Moreover, for the second RG step, only the lower threshold was taken into account, as for data with a contrast agent it is the only meaningful one. While using both thresholds the RG frequently stuck preventing the growth among the whole region. On the other hand, during estimation, it was essential to use both upper and lower thresholds to keep the standard

deviation estimates within reasonable limits. For ARG and ChV the initial mask for the segmentation contained one voxel per each separate region (up to 22 regions) making sure that it is not located on the border of the structure. Parameters for the ChV evolution were adopted using a 10-fold partition and were applied for all data. The same optimization procedure was used to designate ChV parameters, as for $A$ and $B$ in the 2nd step. In some cases, both ARG (3 cases) and ChV (5 cases) lead to leakage to the surrounding structures. These examples were excluded from evaluation of ChV and ARG methods. Regarding the above, the EdgeWave is a stable tool, however it involves manual indication of the segmentation threshold and therefore is prone to subjective setting. Regarding DICE, the Hausdorff distances, both *mHD* and *aHD*, accuracy, sensitivity, under- and over-segmentation rates, the proposed method provided the best results. However, in respect of precision and sensitivity, the EdgeWave algorithm outperformed other evaluated methods. A relatively low performance of ARG and ChV can be attributed to homogeneity assumption of the segmented region which is not true due to improper contrast propagation. The detailed results of the comparison are presented in Table 3 and Fig. 10.

Additionally, the issue of the efficiency of different methods has been addressed. Table 4 presents the overall computation time required by each method (the proposed method, EdgeWave [9], ChV [11] and ARG [8]) regarding the mean calculation time per case. The calculations were performed on a PC equipped with Intel® Core™ i3-3240 CPU, 3.4GHz, 2 cores with 32GB RAM. The input images for each method were derived from the 1st step (ROI designation) and therefore the computation time of this step is excluded from comparison. Such a comparison has several limitations that should be mentioned before making a direct reference to the values presented in Table 4. The ARG was implemented and modified by us and perhaps is different from the original one proposed by Authors regarding its computational efficiency. The mean calculation time required by the EdgeWave was obtained by dividing the total processing time by the number of cases. In addition to the algorithm worktime (approximately 2s per case), also the user-interaction time is taken into account. Moreover, the proposed method, ChV, ARG are Matlab prototypes, whereas the EdgeWave is an executable application. Taking into account the aforementioned issues, interpreting these values should be made with caution. Nevertheless, the EdgeWave was proven to be slightly faster (243.7 s) than the proposed method (250.7 s), whereas the ChV and ARG were much slower (644.3 s and 1047.6 s, respectively).

Table 3. The quantitative results of all evaluated methods; the mean values and the standard deviations are given; the results in bold indicate the best method in each category.
3 examples of ARG and 5 of ChV are excluded.

| | Acc | Prec | Sens | Spec | OS | US | DICE | *mHD* [mm] | *aHD* [mm] |
|---|---|---|---|---|---|---|---|---|---|
| **Proposed method** | **0.997 ±0.004** | 0.872 ±0.107 | **0.883 ±0.074** | 0.998 ±0.003 | **0.166 ±0.165** | **0.141 ±0.101** | **0.871 ±0.060** | **7.73 ±6.67** | **0.46 ±0.36** |
| **EdgeWave** | 0.994 ±0.007 | 0.893 ±0.177 | 0.644 ±0.216 | **0.999 ±0.003** | 0.258 ±0.797 | 2.410 ±8.883 | 0.710 ±0.201 | 18.71 ±11.22 | 2.88 ±2.88 |
| **ChV** | 0.993 ±0.007 | **0.901 ±0.205** | 0.647 ±0.277 | **0.999 ±0.003** | 0.412 ±1.495 | 7.237 ±28.41 | 0.682 ±0.250 | 17.69 ±12.84 | 3.08 ±3.50 |
| **ARG** | 0.985 ±0.020 | 0.703 ±0.323 | 0.684 ±0.684 | 0.988 ±0.022 | 1.128 ±1.809 | 1.191 ±2.813 | 0.583 ±0.218 | 11.82 ±8.16 | 1.99 ±3.25 |

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 3–18.

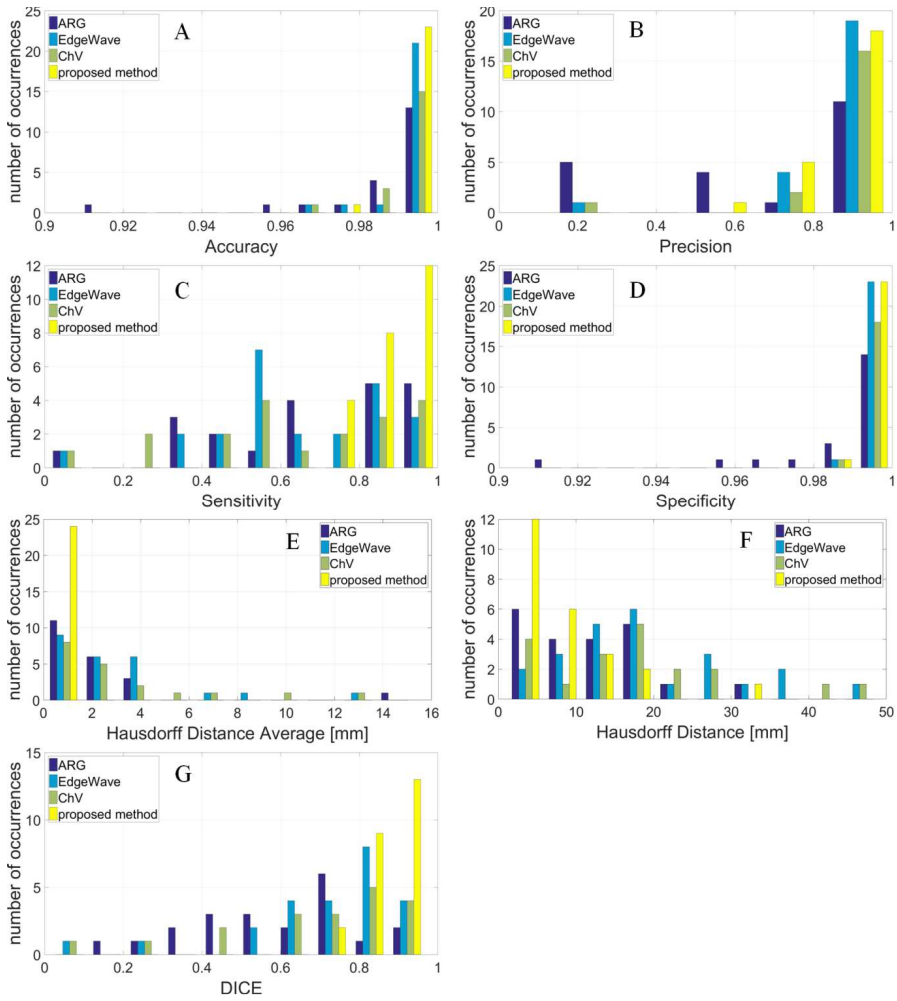



Fig. 10. Compared results of the proposed method, EdgeWave, ChV and ARG in respect of accuracy (A); precision (B); sensitivity (C); specificity (D); *aHD* (E); *mHD* (F) and DICE (G). Due to leakage 3 examples of ARG and 5 of ChV are excluded.

Table 4. The methods' efficiency (computation time) – comparison of the proposed method, EdgeWave [9], ChV [11] and ARG [8] in respect of mean calculation time per case.

| | **Overall Computation Time [s]** | | | |
|---|---|---|---|---|
| **Proposed method** | 2nd step: 0.1 | 3rd step 3 (HLS): 141.6 | 4th step 4 (LARG): 82.1 | **250.7** |
| **EdgeWave** | | | | **243.7** |
| **ChV** | | | | **644.3** |
| **ARG** | | | | **1047.6** |

## 6. Conclusion

This paper proposes a PCS segmentation methodology, consisting of an automatic threshold selection method for binarization (preliminary segmentation), extraction of the PCS parts from

the binarized ROI image by HLS and – finally – the LARG algorithm driven by the image intensity values.

The quantitative evaluation was performed for 24 PCSs belonging to the test set. A significant improvement was observed between the 3rd and 4th steps of the proposed method: the DICE coefficient increased (from 84% to 87%) and the average Hausdorff distance decreased (from 0.74 mm to 0.46 mm). A comparison with the other described methods (ARG, ChV, EdgeWave) revealed that the proposed method outperforms them in respect of DICE coefficient, the maximum and average Hausdorff distances, accuracy, sensitivity, under- and over-segmentation rates. Regarding precision and sensitivity the EdgeWave algorithm provided the best results.

Basing on the proposed PCS segmentations each anatomic detail of the kidney collecting system can be selectively displayed in 3D and used for planning various urological procedures. Additionally, measurements of different objects (the distance between tumour and the PCS, both shape and spatial configuration of individual components of the PCS, the calyceal pattern, the pelvicalyceal angle, a length and width of the lower calyceal neck) are facilitated. The presented PCS segmentation algorithm may constitute a considerable part of a support system for planning minimally invasive procedures. Moreover, the same segmentation framework may be applied to further medical applications. Preoperative imaging and operation planning are the first step in image-guided surgery and now we are planning to study this approach on clinical aspects, especially on a complication rate during PN, PCNL or RIRS. The next step will be development of intraoperative imaging and tracking systems.

## Acknowledgements

The work was supported by the Ministry of Science and Higher Education, Poland (statutory activity no. 11.11.120.774, Dean Grant no. 15.11.120.889).

We would like to thank prof. Henry Rusinek for sharing the application FireVoxel containing the EdgeWave algorithm (*https://wp.nyu.edu/firevoxel/*) and valuable comments facilitating its handling.

## References

[1] *Statistical Bulletin of the Ministry of Health*. (2016). Centre for Health Information Systems.

[2] Ukimura, O., *et al*. (2012). Three-dimensional reconstruction of renovascular-tumor anatomy to facilitate zero-ischemia partial nephrectomy. *European Urology*, 61(1), 211–217.

[3] Ljungberg, B., *et al*. (2015). Guidelines on Renal Cell Carcinoma. *European Association of Urology*.

[4] Zöllner, F.G., *et al*. (2012). Assessment of Kidney Volumes From MRI: Acquisition and Segmentation Techniques. *American Journal of Roentgenology,* 199 (5), 1060–1069.

[5] Will, S., *et al*. (2014). Automated segmentation and volumetric analysis of renal cortex, medulla, and pelvis based on non-contrast-enhanced T1-and T2-weighted MR images. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 27(5), 445–454.

[6] Yang, X., *et al*. (2015). Automatic Segmentation of Renal Compartments in DCE-MRI Images. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*, 3–11.

[7] Li, X., *et al*. (2011). Renal cortex segmentation using optimal surface search with novel graph construction. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2011,* 387–394.

[8] Pohle, R., Toennies, K.D. (2001). A new approach for model-based adaptive region growing in medical image analysis. *Computer Analysis of Images and Patterns Journal*, 238–246.

[9] Abiria, B., *et al.* (2014). Performance of an automated renal segmentation algorithm based on morphological erosion and connectivity. *Proc. of SPIE, Medical Imaging 2014: Computer-Aided Diagnosis*, 90352R–90352R–5.

[10] Sun, Y. *et al.* (2002). Kidney segmentation in MRI sequences using temporal dynamics. *Proc. of IEEE International Symposium on Biomedical Imaging*, 98–101.

[11] Chan, T.F., Vese, L.A. (2001). Active contours without edges. *IEEE Trans. on image processing*, 10(2), 266–277.

[12] Shehata, M., *et al.* (2015). A level set-based framework for 3D kidney segmentation from diffusion MR images. *Proc. of International Conference on Image Processing*, 4441–4445.

[13] Song, T., *et al.* (2008). Segmentation of 4D MR Renography Images Using Temporal Dynamics in a Level Set Framework. *Proc. of IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 37–40.

[14] Bugajska, K. *et al.* (2015). The renal vessel segmentation for facilitation of partial nephrectomy. *Proc. of IEEE, SPA: Signal Processing: Algorithms, Architectures, Arrangements and Applications*, 50–55.

[15] Yushkevich, P.A., *et al.* (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3), 1116–1128.

[16] Pankaj, R., *et al.* (2014.) Radiological Study of Variations in the pelvicalyceal system of kidney. *International Journal of Biological and Medical Research*, 5(3), 4336–4339.

[17] Sampaio, F.J.B., *et al.* (1988). Anatomic classification of the kidney collecting system for endourologic procedures. *Endourology,* 2(3), 247–251.

[18] Sampaio, F.J.B., *et al.* (1992). Inferior pole collecting system anatomy: its probable role in extracorporeal shock wave lithotripsy. *Journal of Urology*, 147, 322–324.

[19] Keeley, F.X. Jr., *et al.* (1999). Clearance of lower-pole stones following shock wave lithotripsy: effect of the infundibulopelvic angle. *European Urology*, 36, 371–375.

[20] Sumino, Y., *et al.* (2002). Predictors of lower pole renal stone clearance after extracorporeal shock wave lithotripsy. *Journal of Urology*, 168, 1344–1347.

[21] Lin, C.C., *et al.* (2008). Predictive factors of lower calyceal stone clearance after extracorporeal shockwave lithotripsy (ESWL): the impact of radiological anatomy. *Journal of the Chinese Medical Association,* 71(10), 496–501.

[22] Resorlu, B., *et al.* (2012). The Impact of Pelvicaliceal Anatomy on the Success of Retrograde Intrarenal Surgery in Patients With Lower Pole Renal Stones. *Urology*, 79(1), 61–66.

[23] Elbahnasy, A.M., *et al.* (1998). Lower caliceal stone clearance after shock wave lithotripsy or ureteroscopy: the impact of lower pole radiographic anatomy. *Journal of Urology*, 159, 676–682.

[24] Fong, Y.K., *et al.* (2004). Lower pole ratio: a new and accurate predictor of lower pole stone clearance after shockwave lithotripsy. *International Journal of Urology*, 11, 700–703.

[25] Türk, C., *et al.* (2015). Guidelines on Urolithiasis. *European Association of Urology*.

[26] Xu, Y., *et al.* (2016). The value of three-dimensional helical computed tomography for the retrograde flexible ureteronephroscopy in the treatment of lower pole calyx stones. *Chronic Diseases and Translational Medicine*, available online Apr. 06, 2016, DOI:10.1016/j.cdtm.2016.02.001.

[27] Sargin, S.Y., *et al.* (2014). The efficacy of radiographic anatomical measurement methods in predicting success after extracorporeal shockwave lithotripsy for lower pole kidney stones. *International Brazilian Journal Of Urology*, 40(3), 337–345.

[28] Sonka, M., *et al.* (2008). *Image processing, analysis, and machine vision*. 3rd ed. Cengage Learning.

[29] Byrd, R.H., *et al.* (1988). Approximate Solution of the Trust Region Problem by Minimization over Two Dimensional Subspaces. *Mathematical Programming*, 40, 247–263.

[30] Lagarias, J.C., *et al.* (1998). Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, 9(1), 112–147.

[31] Zhang, Y., *et al*. (2008). Medical image segmentation using new hybrid level-set method. *Proc. of BioMedical Visualization MEDIVIS'08 IEEE*, 71–76.

[32] Skalski, A., *et al*. (2017). Kidney segmentation in CT data using hybrid Level-Set Method with ellipsoidal shape constraints. *Metrol. Meas. Syst.*, 24(1), 101–112.

[33] Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302.

[34] Huttenlocher, D.P., *et al*. (1993). Comparing images using the Hausdorff distance. *IEEE Tran. on Pattern Analysis and Machine Intelligence,* 15(9), 850–863.

[35] Dubuisson, M.P., *et al*. (1994). A modified Hausdorff distance for object matching. *Proc. of International Conference on Pattern Recognition*, 566–568.

[F1] *https://en.wikipedia.org/wiki/Kidney#/media/File:KidneyStructures_PioM.svg.*

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 19–26.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# MEASUREMENTS OF CONCENTRATION DIFFERENCES BETWEEN LIQUID MIXTURES USING DIGITAL HOLOGRAPHIC INTERFEROMETRY

**Carlos Guerrero-Méndez[1)], Tonatiuh Saucedo-Anaya[2)], Maria Araiza-Esquivel[1)], Raúl E. Balderas-Navarro[3)], Alfonso López-Martínez[1)], Carlos Olvera-Olvera[1)]**

*1) Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Ramón López Velarde 801, C.P. 98000, Zacatecas, México (✉ capacti@gmail.com, +52 492 92 29 699, araizamae@yahoo.com, alopez2601@hotmail.com, olveraca@gmail.com)*

*2) Universidad Autónoma de Zacatecas, Unidad Académica de Física, Calzada Solidaridad Esq. Con Paseo La Bufa S/N, C.P. 98060, Zacatecas, México (tsaucedo@fisica.uaz.edu.mx)*

*3) Instituto de Investigación en Comunicación Óptica (IICO-UASLP), Karakorum 1470, Lomas 4ta. Sección, C.P. 78210 , San Luis Potosí, México (raul.balderas@gmail.com)*

**Abstract**

We present an alternative method to detect and measure the concentration changes in liquid solutions. The method uses *Digital Holographic Interferometry* (DHI) and is based on measuring refractive index variations. The first hologram is recorded when a wavefront from light comes across an ordinary cylindrical glass container filled with a liquid solution. The second hologram is recorded after slight changing the liquid's concentration. Differences in phase obtained from the correlation of the first hologram with the second one provide information about the refractive index variation, which is directly related to the changes in physical properties related to the concentration. The method can be used – with high sensitivity, accuracy, and speed – either to detect adulterations or to measure a slight change of concentration in the order of 0.001 moles which is equivalent to a difference of 0.003 g of sodium chloride in solutions. The method also enables to measure and calculate the phase difference among each pixel of two samples. This makes it possible to generate a global measurement of the phase difference of the entire sensed region.

Keywords: Digital Holographic Interferometry, refractive index measurements, phase difference, full-field measurements.



## 1. Introduction

Liquid mixtures can be classified based on their physical properties such as concentration, weight, colour, and boiling temperature, among others [1]. The concentration of a liquid solution refers to the amount of solute (in moles or mass) dissolved in a certain quantity of solvent [2]. Methods and tools for accurate measurements that can detect slight concentration variations are greatly important for science, regulatory agencies, food processors, and consumers. Expensive liquids, including olive oil, fruit juices, honey, alcoholic drinks, and gasoline, are especially vulnerable to adulteration. For this reason, a fast and accurate technique is required to validate the concentrations of products or liquid mixtures. Optical techniques are non-destructive and are generally preferred for this purpose.

The index of refraction is one of the most important optical properties of an object [3]. In liquid solutions, this parameter is unique and proportional to the concentration of a substance [4]. Commonly, the refractive index is determined using Snell's law, which involves the displacement of the angle of an incident beam with respect to a refracted beam by a phase object. Some methods based on this law use prisms [5–8], squares [9, 10], and special containers [11]. However, these methods require a good estimation of the angles, which reduces their accuracy. Other disadvantages are that they use only a small region (scarcely a point) to obtain

Article history: received on Dec. 03, 2015; accepted on Aug. 28, 2016; available online on Mar., 10, 2017; DOI: 10.1515/mms-2017-0002.

the refractive index of a sample, and the systems are difficult to calibrate and apply in real environments.

New full-field optical techniques have been developed that are more precise, accurate, non-destructive, and non-invasive. These methods have high resolution and stability, and they can measure profiles of physical variations in mixtures [12–14]. The traditional techniques that have been used to measure and visualize refractive index variations are the Schlieren, shadowgraph, and interferometry techniques, from which *Digital Holographic Interferometry* (DHI) has been developed [15]. Important efforts have been made to establish refractive index values using DHI [16]. They are related to concentration variations in liquid samples [17–18]. However, these methods use a special container and require knowledge of the dimensions of the container in advance. Also, they provide point measurements and are not able to take global measurements of a sample.

We present a fast, simple, high-precision, non-destructive, full-field optical technique for measuring concentration differences between liquid mixtures. The proposed method can obtain information from every small region of the wavefront coming from each sample being analyzed. All the regions are then used to calculate the global variation using the concentration variations of the samples. The process of phase retrieval is carried out digitally using the Fourier method [19]. This method uses an ordinary cylindrical container, which makes its implementation easier for industrial processes. Commonly, tubes are used to transport liquid products, and the proposed method makes it possible to monitor the concentration of liquid products during transport.

The remaining of the paper is organized as follows: in Section 2, we explain operation of the proposed optical system. Section 3 presents the numerical principles, the phase estimation method and the relation between a phase difference and a concentration variation of two liquid solutions. The experimental results are reported in Section 4. Finally, in Section 5, we summarize the conclusions of our work.

## 2. Experimental setup

A schematic diagram for detecting and measuring the concentration changes using DHI is shown in Fig. 1.

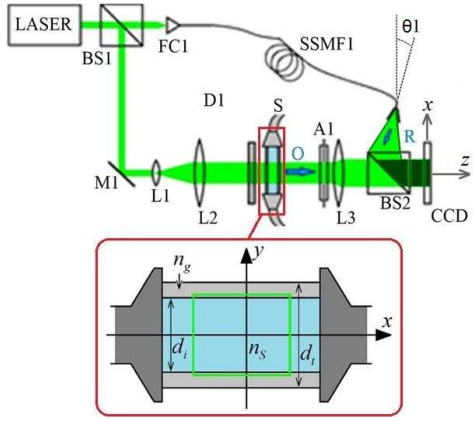



Fig. 1. A schematic diagram of the experimental setup using DHI. BS1, BS2: cubic beam splitters; FC1– a fibre collimator; M1 – a mirror; L1, L2, L3 – lenses; SSMF1 – a single-mode fibre; S – a liquid sample; D1 – a diffuser; A1 – an aperture; O – an object beam; R – a reference beam; $\theta 1$ represents the carrier spatial frequency along the direction $x$ of the sensor plane. The wavefront comes from the green region in the glass view; $x'$ and $y'$are the rectangular coordinates of the container with the liquid inside.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 19–26.

Monochromatic He-Ne laser light with $\lambda = 543$ nm and a maximum output power of 15 mW is split into two beams by a beam splitter BS1. The reflected beam (the "object beam") from mirror M1 is reflected towards lenses L1 and L2 (expanded and collimated ones, respectively) and a diffuser D1. The beam passes through an ordinary glass tube with unknown inner dimensions $d_i$ and containing a liquid sample S. This object beam enters through a rectangular aperture A1 and is collected by a positive lens L3, which creates on a CCD sensor an image of the tube containing the sample. The transmitted beam (the "reference beam") travels through a single-mode optical fibre SSMF1. It is sent into a cubic beam splitter BS2, which is placed in front of the CCD in such a way that it interferes with the object beam. Thus, a hologram ($H_S$) is recorded from the aqueous sample. The liquid solutions to be analysed are injected into the tube at a constant rate (~36 ml/s), and the interference patterns are recorded using a CCD, which is a monochromatic sensor with 1280 × 1024 pixels (1.3 MP) and a pixel size of 6.7 µm × 6.7 µm. All digital processing is done using Matlab. When recording the holograms, the temperature was stabilized at 20°C.

## 3. Method

The holographic technique can record the amplitude and phase (complete information) of a wave-front scattered by an object. The holographic interferometry setup uses the holography method to interferometrically compare two or more wave-fronts recorded at different moments or states [14]. The results of the comparison are used to obtain the phase difference map, which shows the physical variations between two liquids.

In order to measure the concentration difference between two liquid mixtures, we recorded two holograms that describe the substance coming from each liquid sample. By using the DHI double exposure method and an ordinary glass tube as an object, we obtained a hologram $H_{S_1}$ from a wave-front coming from the tube filled with a certain liquid solution $S_1$ in the optical system (see Fig. 1). This can be represented using:

$$U_{S_1} = u_{S_1}(x,y)\ \exp[i\phi_{S_1}(x,y)], \tag{1}$$

where: $u_{S_1}$ represents the amplitude; $\varphi_{S_1}$ is the phase of the wavefront; and $x$ and $y$ are rectangular coordinates of the recording sensor plane. A second hologram $H_{S_2}$ is then recorded either using another liquid solution or after slightly modifying the concentration of the liquid sample (creating $S_2$). The new phase is $\varphi_{S_2}$, which indicates a change in the optical path length. $\varphi_{S_2} = \varphi_{S_1} + \Delta\varphi_{S_2-S_1}$, which creates a wavefront that can be expressed as:

$$U_{S_2} = u_{S_2}(x,y)\ \exp\{i[\phi_{S_1}(x,y) + \Delta\phi_{S_2\text{-}S_1}(x,y)]\}, \tag{2}$$

or simply:

$$U_{S_2} = u_{S_2}(x,y)\ \exp[i\phi_{S_2}(x,y)]. \tag{3}$$

The two wave-fronts scattered by the tube have a phase distribution due to the morphological and physical properties of the object phase (see the red part of Fig. 1). The phase of the wave-fronts can be represented as:

$$\phi_m = k\{[d_t(x,y) - d_i(x,y)]n_g(x,y) + d_i(x,y)n_{S_m}(x,y)\}, \quad m = 1,2, \tag{4}$$

where $k = 2\pi/\lambda$; $d_i$ and $d_t$ are the inner and outer transversal distances of the glass tube; $n_{S_m}$ and $n_g$ are the refractive indices of the mixture and the glass walls, respectively.

### *3.1. Phase measurement*

The total intensity recorded on the electronic sensor using any liquid sample in the tube is expressed by:

$$I(x,y) = |R(x,y)|^2 + |U(x,y)|^2 + U(x,y)R^*(x,y) + R(x,y)U^*(x,y), \tag{5}$$

where $U(x,y) = u(x,y)\exp[i\varphi(x,y)]$ and $R(x,y) = r(x,y)\exp[-i2\pi(f_x x + f_y y)]$, which are the complex amplitudes of the liquid mixture and the reference beam, respectively. $f_x = (\sin\theta 1)/\lambda$ and $f_y = (\sin\theta 2)/\lambda$ create a spatial frequency along the $x$ and $y$ directions caused by a small inclination $\theta 1$ and $\theta 2$ of the reference beam, since only the phase of the reference beam changes according to the register media, and "$*$" denotes the complex conjugate.

Equation (5) can be written as:

$$I(x,y) = a(x,y) + c(x,y)\exp[i2\pi(f_x x + f_y y)] + c^*(x,y)\exp[-i2\pi(f_x x + f_y y)], \tag{6}$$

where $a(x,y) = u^2(x,y) + r^2(x,y)$ and $c(x,y) = r(x,y)u(x,y)\exp[i\varphi(x,y)]$.

The size of the aperture was chosen in order to obtain a greater amount of high frequencies in the Fourier spectrum. In order to obtain the phase term in every hologram, a Fourier transform must be performed on (6), which is expressed as:

$$FT\{I(x,y)\} = A(\mu,\nu) + C(\mu - f_x, \nu - f_y) + C^*(\mu + f_x, \nu - f_y), \tag{7}$$

where capital letters represent the Fourier transform (see Fig. 2), while $(\mu,\nu)$ are the spatial frequencies in the $x$ and $y$ directions, respectively.

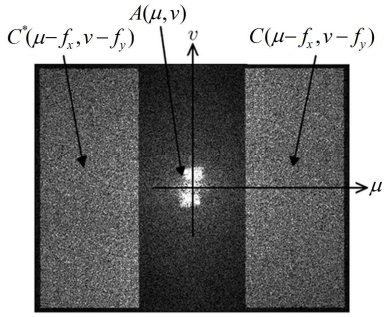



Fig. 2. A Fourier spectrum with the aperture.

The complex conjugate terms $C$ or $C^*$ are used to obtain the required phase term of the reconstructed wave-fronts. From this, only one of the three terms is filtered. Its inverse Fourier transform is then calculated to obtain the phase distribution:

$$\varphi(x,y) + 2\pi(f_x x + f_y y) = \arctan\frac{\mathrm{Im}[c(x,y)]}{\mathrm{Re}[c(x,y)]}. \tag{8}$$

The complete phase recovery process is visualized in Fig. 3 and can also be seen in previous studies [19, 21].

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 19–26.

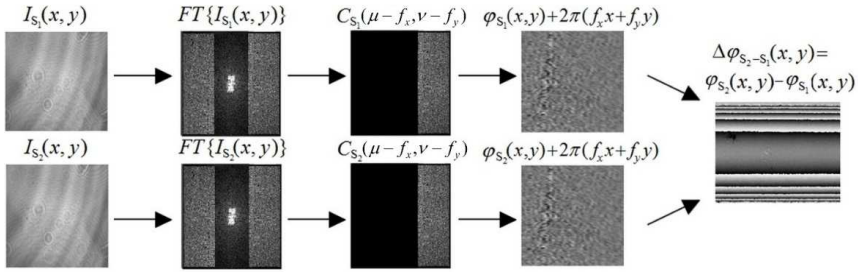



Fig. 3. The phase recovery process.

### *3.2. Concentration difference in liquid*

With the individual phase terms $H_{S_1}$ and $H_{S_2}$, the procedure continues with the calculation of the phase difference $\Delta\varphi_{S_2-S_1} = \varphi_{S_2} - \varphi_{S_1}$. A phase term depends on the transverse distances and the refractive index of the liquid mixture inside a glass tube. Thus, we can represent this phase difference as:

$$\Delta\phi_{S_2-S_1}(x,y) = k\{d_i(x,y)[\Delta n_{S_2-S_1}(x,y)]\}, \tag{9}$$

where $\Delta n_{S_2-S_1}(x,y)$ is the refractive index difference between substances $S_2$ and $S_1$.

The refractive index difference is related to the change of concentration $CON$ and the temperature $T$ between substances. Then, $\Delta n_{S_2-S_1}$ in (9) can be expressed as:

$$\Delta n_{S_2-S_1}(x,y) = \left[\frac{\partial n_S}{\partial CON}\right]_T [CON_{S_2}(x,y) - CON_{S_1}(x,y)] + \left[\frac{\partial n_S}{\partial T}\right]_{CON} [T_{S_2}(x,y) - T_{S_1}(x,y)], \tag{10}$$

where $\left[\frac{\partial n_S}{\partial CON}\right]_T$ and $\left[\frac{\partial n_S}{\partial T}\right]_{CON}$ are values that represent the dependence of the refractive index on $CON$ and $T$, respectively. $CON_{S_2}$ and $T_{S_2}$ are the concentration and the temperature of $S_2$, whereas $CON_{S_1}$ and $T_{S_1}$ are those of $S_1$.

Aqueous salt mixtures (NaCl + $H_2O$) have a linear relationship between $n$ and $CON$ $\left(\left[\frac{\partial n_S}{\partial CON}\right]_T\right)$, which is considered to be constant at $1.71 \times 10^{-3}$ at a temperature of 20°C. Then, (9) can be written as:

$$\Delta\phi_{S_2-S_1}(x,y) = k\{d_i(x,y)[1.71\times10^{-3}][CON_{S_2}(x,y) - CON_{S_1}(x,y)]\}, \tag{11}$$

Using (11), we can calculate the concentration difference between $S_1$ and $S_2$, but $d_i$ is not known because we used an ordinary glass cylinder whose walls are optically imperfect. To solve this issue, we used a reference solution $S_{H_2O}$ and another liquid mixture ($S_{H_2O+NaCl}$) with known parameters to create an independent expression that eliminates the dependence on $d_i$. Then, we need to create another phase difference $\Delta\varphi_{ref}$ using these two solutions. We employed it to obtain $d_i$ as:

$$d_i(x,y)=\frac{\Delta\phi_{ref}(x,y)}{[1.71\times10^{-3}][\Delta CON_{S_{ref}}(x,y)]}k^{-1}, \tag{12}$$

where $\Delta\varphi_{ref}=\varphi_{H_2O+NaCl}-\varphi_{H_2O}$ and $\Delta CON_{ref}=CON_{H_2O+NaCl}-CON_{H_2O}$.

Using (11) and (12), we can calculate the concentration difference between two substances as:

$$\Delta CON_{S_2-S_1}(x,y)=\frac{\Delta\phi_{S_2-S_1}(x,y)}{\Delta\phi_{ref}(x,y)}\left[\Delta CON_{ref}(x,y)\right]. \tag{13}$$

By employing (13), the full-field distribution of the concentration difference in a glass tube can be calculated and visualized.

## 4. Results

In order to verify operation of the experimental setup, we calculated and visualized the global concentration difference distribution between saline mixtures. The liquid samples were prepared by mixing distilled water ($S_{H_2O}$) (50 ml) and definite quantities of NaCl (0.25, 0.5, 0.75, 1, 1.25 and 1.5 g) to create each mixed solution (NaCl + $H_2O$) with specific molarities of $S_{mol_1}=0.086$, $S_{mol_2}=0.172$, $S_{mol_3}=0.258$, $S_{mol_4}=0.344$, $S_{mol_5}=0.43$, and $S_{mol_6}=0.516$ moles. A set of $\Delta\varphi_{S_2-S_1}$ was calculated for solutions with concentration differences of 0.086 mol between them. The first solution with a lower concentration is taken as $S_1$, and the next liquid solution with a higher concentration – as $S_2$ (see Figs. 4a–4f).

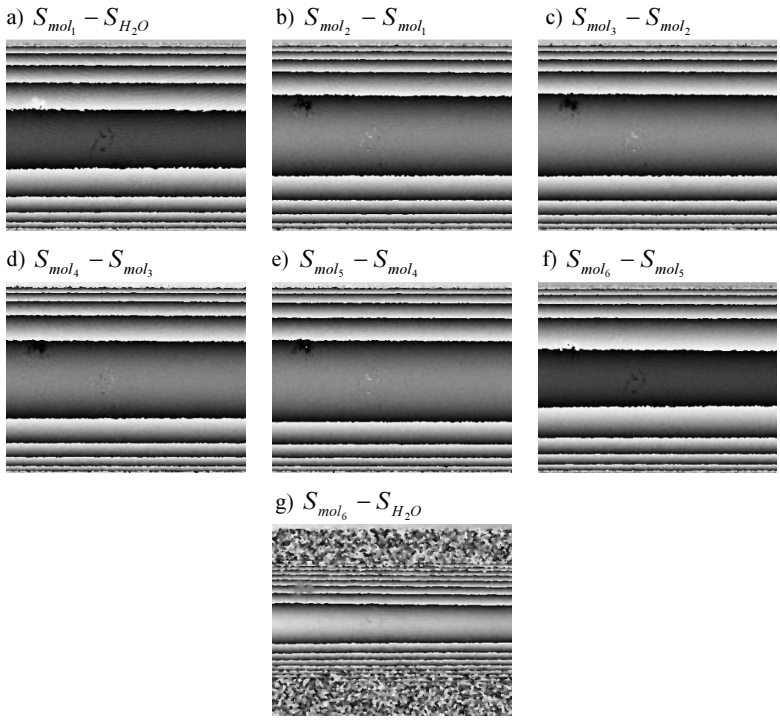



Fig. 4. Wrapped phase difference maps.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 19–26.

$\Delta\varphi_{ref}$ was calculated in all the experiments using $S_{\mathrm{H_2O}}$ and $S_{mol_1}$ as $S_{\mathrm{H_2O+NaCl}}$. For example, to calculate the concentration difference between $S_{mol_5}$ and $S_{mol_6}$, we took the first one as $S_1$ and the second (with a higher molarity) as $S_2$ to create $\Delta\phi_{S_2-S_1}$, together with the other values ($\Delta\varphi_{ref}$ and $\Delta CON_{ref}$). Then, the distribution of the concentration difference between the mixtures can be calculated using (13). The values of $CON$ and $n$ of these last two solutions were obtained from [1] and [20]. The concentration difference values obtained with the proposed method are presented in Table 1.

Table 1. Comparisons of concentration values measured by the DHI and those found in [20].

| Solutions compared | $\Delta CON_{S_2-S_1}$ (value in Ref. 20) [*mol*] | $\Delta CON_{S_2-S_1}$ (with DHI) [*mol*] | Deviation |
|---|---|---|---|
| $S_{mol_1} - S_{H_2O}$ | 0.083 | 0.083 | 0.0 |
| $S_{mol_2} - S_{mol_1}$ | 0.086 | 0.086 | 0.0 |
| $S_{mol_3} - S_{mol_2}$ | 0.086 | 0.088 | +0.002 |
| $S_{mol_4} - S_{mol_3}$ | 0.086 | 0.081 | –0.005 |
| $S_{mol_5} - S_{mol_4}$ | 0.086 | 0.082 | –0.004 |
| $S_{mol_6} - S_{mol_5}$ | 0.086 | 0.085 | –0.001 |

The performance of the CCD sensor employed in this experiment was assessed using liquid substances with concentration differences of 0.086 moles between them. Substances with a higher concentration difference generate a wrapped phase map with a high frequency, which is more difficult to unwrap and does not enable to obtain the concentration differences. For example, if we generate a phase difference between the liquid samples of $S_{\mathrm{H_2O}}$ and $S_{mol_6}$, we obtain the phase difference map shown in Fig. 4g.

## 5. Conclusions

This work has presented a method of detecting and measuring the global concentration variations in liquid mixtures using DHI. The process measures phase variations between wavefronts scattered by an ordinary glass tube and converts them with a phase change into a concentration variation. The method is non-invasive, simple, fast, and easy to develop in a laboratory and real work environments. The technique can resolve extremely small changes of concentration in the order of –0.001 moles, which is equivalent to a difference of 0.003 g of sodium chloride in saline solutions. In other words, since we used 50 ml of distilled water, the method can distinguish changes in salt concentration of $6 \times 10^{-5}$ by weight. Additionally, the method does not require a special device to contain the saline sample. The results are in accordance with concentration values published in [20] on aqueous salt solutions ($NaCl + H_2O$). Our method can be used to identify or confirm the identity of a sample, as well as to detect adulterations or fake solutions.

## Acknowledgments

One of the authors (Carlos Guerrero-Méndez) acknowledges CONACYT (México) for providing a partial financial support for this work.

## References

[1] Cracolice, M.S. (2016). *Basics of Introductory Chemistry with Math Review*. Montana: Brooks/Cole.

[2] Henrickson, C. (2010). *CliffsNotes Chemistry Practice Pack*. Ney Jersey: J. Wiley & Sons.

[3] Hecht, E. (2002). *Optics.* 4th ed. San Francisco: Addison-Wesley.

[4] Kress-Rogers, E., Brimelow, C.J.B. (2001). *Instrumentation and Sensors for the Food Industry.* 2nd ed. Abington: Woodhead Pubishing Limited.

[5] Chandra, B., Bhaiya, S. (1983). A simple, accurate alternative to the minimum deviation method of determining the refractive index of liquids. *Am. J. Phys.*, 51(2), 160–161.

[6] Grange, B., Stevenson, W.H., Viskanta, R. (1976). Refractive index of liquid solutions at low temperatures: an accurate measurement. *Appl. Opt.*, 15(4), 858–859.

[7] Edmiston, M.D. (1986). Measuring refractive indices. *Phys. Teach.*, 24(3), 160–163.

[8] Shenoy, M.R.S., Thyagarajan, K. (1990). Simple prism coupling technique to measure the refractive index of a liquid and its variation with temperature. *Rev. Sci. Instrum.*, 61(3), 1010–1013.

[9] Fan, J.P.L.C-H. (1998). Precision laser-based concentration and refractive index measurement of liquids. *Microscale Thermophysical Engineering*, 2(4), 261–272.

[10] Nemoto, S. (1992). Measurement of the refractive index of liquid using laser beam displacement. *Appl. Opt.*, 31(31), 6690–6694.

[11] Moreels, E., De, Greef C., Finsy, R. (1984). Laser light refractometer. *Appl. Opt.*, 23(17), 3010–3013.

[12] Toker, G.R. (2012). *Holographic interferometry: A Mach-Zehnder Approach.* Boca Raton: Taylor & Francis Group.

[13] Colombani, J., Bert, J. (2007). Holographic interferometry for the study of liquids. *J. Mol. Liq.*, 134(1), 8–14.

[14] Kreis, T. (2005). *Handbook of holographic interferometry: Optical and Digital Methods*. Klagenfurter: WILEY-VCH Verlag GmbH & Co.KGaA.

[15] Goldstein, R.J. (1996). *Fluid mechanics measurements.* 2nd ed. Philadelphia: Taylor & Francis Group.

[16] Hossain, M.M., Mehta, D.S., Shakher, C. (2006). Refractive index determination: an application of lensless fourier digital holography. *Opt. Eng.*, 45(10), 106203–106203.

[17] Zhang, Y., Zhao, J., Di J., Jiang, H., Wang, Q., Wang, J., Guo, Y., Yin, D. (2012). Real-time monitoring of the solution concentration variation during the crystallization process of protein-lysozyme by using digital holographic interferometry. *Opt. Express*, 20(16), 18415–18421.

[18] Zhao, J., Zhang, Y., Jiang, H., Di, J. (2013). Dynamic measurement for the solution concentration variation using digital holographic interferometry and discussion for the measuring accuracy. *Proc. icOPEN2013*, Singapore, Singapure, 87690D–87690D.

[19] Takeda, M., Ina, H. Kobayashi, S. (1982). Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry. *Jos. A.*, 72(1), 156–160.

[20] Haynes. W.M. (2015). *Concentrative properties of aqueous solutions: density, refractive index, freezing point depression, and viscosity 96th ed.*, Boca Raton: Taylor & Francis Group.

[21] Saucedo, A.T., Mendoza, F., De la Torre-Ibarra M., Pedrini, G., Osten, W. (2006). Endoscopic pulsed digital holography for 3D measurements. *Opt. Express*, 14(4), 1468–1475.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# MINING DATA OF NOISY SIGNAL PATTERNS IN RECOGNITION OF GASOLINE BIO-BASED ADDITIVES USING ELECTRONIC NOSE

**Stanisław Osowski[1,2], Krzysztof Siwek[1]**

*1) Warsaw University of Technology, Faculty of Electrical Engineering, Pl. Politechniki 1, 00-661 Warsaw, Poland (✉ sto@iem.pw.edu.pl, +48 22 234 7235, ksiwek@iem.pw.edu.pl)*

*2) Military University of Technology, Faculty of Electronic Engineering, Gen. S. Kaliskiego 1, 00-908 Warsaw, Poland*

**Abstract**

The paper analyses the distorted data of an electronic nose in recognizing the gasoline bio-based additives. Different tools of data mining, such as the methods of data clustering, principal component analysis, wavelet transformation, support vector machine and random forest of decision trees are applied. A special stress is put on the robustness of signal processing systems to the noise distorting the registered sensor signals. A special de-noising procedure based on application of discrete wavelet transformation has been proposed. This procedure enables to reduce the error rate of recognition in a significant way. The numerical results of experiments devoted to the recognition of different blends of gasoline have shown the superiority of support vector machine in a noisy environment of measurement.

Keywords: data mining, electronic nose, gasoline blends, random forest, support vector machine, wavelet de-noising.



## 1. Introduction

The subject of the paper is studying the properties of sensor signals distorted by the random noise in an electronic nose. The noise present in the nose measurement may be a result of thermal fluctuations, power drift, sensor instability or chemical interfering agents, *etc*. The additive Gaussian model of noise affecting the semiconductor sensors is usually assumed in considerations and this model has been examined in the paper. The noise makes measurements not-repeatable, so analyzing the effect of this distortion is significant from a practical point of view.

In the paper this distortion is statistically analyzed on the examples of gasoline blends built on the basis of different bio-products, such as ethanol, *methyl tertiary butyl ether* (MTBE*), ethyl tertiary butyl ether* (ETBE) and benzene. These supplements are used as fuel oxygenates to increase the octane index (to replace the banned tetraethyl lead) or to raise the oxygen content in gasoline. The addition of the bio-products changes the odor of a blend and thus enables to recognize its type [1, 2]. An electronic nose, being able to recognize different types of blends, may find practical application in building a low-cost, very fast and accurate measuring instrument for recognizing the gasoline blends. This is an important problem in checking the quality and parameters of gasoline blends sold in petrol stations.

Different types of sensors are used in an electronic nose. A good review of them is presented in the papers [3] and [4]. In our application we have used an array of semiconductor sensors, very popular due to their wide availability, ease of use and relatively low cost. They form the heart of an electronic nose and respond with a signal pattern according to the odor of each gasoline blend. Analysis of these signals can answer questions regarding recognition of the supplemented bio-products directly on the basis of the blend odor [2, 3, 5]. One of the most

Article history: received on Aug. 11, 2016; accepted on Jan. 14, 2017; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0015.

important problems in this task is assuring the robustness of a nose to the disturbed measurement signals in connection with the number of applied sensors. The distortion of sensor signals is recognized as small and non-deterministic temporal variations. It is the result of changing the working conditions of the semiconductor sensors following the changes of the environmental parameters (temperature, humidity, pressure), as well as sensor circuit instability, which occur in the measurement process [6, 7].

There are some papers studying the effect of noise and drift in the nose measurement process and methods of its reduction [6–10]. They are based on application of compression techniques and apply the methods of principal component analysis, wavelet transformation or singular value decomposition. Most of the papers concerning the pattern recognition in an electronic nose deal only with the actually measured signals, not analyzing the effect of their additional distortion by noise [1–3]. The most often used techniques apply the neural networks, the support vector machine, the principal component analysis, the linear discriminant analysis, *etc.*

The aim of the paper is to study the statistical effect of noise disturbing the semiconductor sensor signals in an electronic nose, used in the process of recognition of the gasoline blends.

The problem of a long time drift is not considered. Different levels of white Gaussian noise are applied and a few aspects of the problem are considered:

- The unsupervised analysis of data distribution at different levels of distortion; it is based on clustering the measured samples and analyzing their changes caused by noise.
- The supervised analysis directed to pattern recognition and classification, in order to determine the robustness of system to noise.
- Improvement of the pattern recognition accuracy by applying a de-noising procedure to the noisy sensor signals.

The paper compares the performance of two most efficient classification systems: the random forest of decision trees [11] and the support vector machine [12]. The random forest proposed by Breiman in 2001 is based on a learning strategy called "ensemble learning" with generating many classifiers and aggregating their results according to the majority voting rule. The random forest can be directly applied to solve the recognition and classification tasks. It also provides a measure of significance of a particular sensor in the measurement process. In this respect it is a very useful tool, more universal than most of the known solutions of signal processing in an electronic nose, such as the linear discriminant analysis, *principal component analysis* (PCA), neuro-fuzzy systems or neural networks [13–17]. On the other hand, the support vector machine was created as a classification tool significantly resistant to noise distorting the input data [12, 17, 18]. Both methods are compared for the data distorted by the random noise with a normal distribution, zero mean and different variance values.

An important aspect considered in the paper is reduction of the noise contaminating the sensor signals. The discrete wavelet transform, decomposing sensor signal patterns into wavelet components of different resolution, is studied. The noise influence is reduced by cutting the detailed signals of their lowest values. Thanks to this the final classification accuracy of patterns is achieved. The results of numerical experiments have shown a high efficiency of this strategy in improving the recognition of patterns formed by the sensor signals.

## 2. The electronic nose measurements

Recognition of the gasoline blends on the basis of their odor exploits the fact that the blends are associated with different odors resulting from their chemical composition. Analysis of the odor is a complex issue because of a heterogeneous nature of gasoline. Application of an electronic nose and artificial intelligence methods seems to be an efficient way of analyzing the odor [2–4]. The patterns of signals of the vapor-sensitive detectors are processed in this approach and associated with different types (classes) of gasoline blends.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

Many different techniques of signal processing have been employed in an electronic nose. They include: the *principal component analysis* (PCA) and linear discriminant analysis [3], self-organizing maps [16], the *k*-nearest neighbor algorithm [15], neuro-fuzzy systems [14], different types of neural networks [1, 3], the support vector machine [2, 17, 18]. Recently, the random forest has been also tried in recognition of orange beverage and Chinese vinegar [18]. Application of the random forest seems to be an interesting alternative to the most often used neural networks.

An array of seven tin oxide-based gas sensors (TGS815, TGS821, TGS822, TGS825, TGS824, TGS842 and TGS822 modified by using an additional resistive potentiometer circuit) from Figaro Engineering Inc., mounted into an optimized test chamber has been applied in the computerized measurement system [2] (shape: cylinder, response time of sensors: 120 s, volume 0.2l of the test chamber was adjusted to the flow rate and time response). The carrier gas (synthetic air) flows through the chamber in controlled temperature conditions. The capacity of the measurement chamber, the carrier flow, the temperature and the size of gasoline sample are kept constant during the measurements. The signals are acquired by using an ADAM-4017 type 8-Channel Analog Input Module Rev.D1 and a serial communication interface with a PC computer.

The diagnostic features have been extracted from the averaged temporal series of sensor resistances $R(j)$, one for each $j$-th sensor ($j = 1, 2, \ldots, 7$) of the array. They are defined as relative variations $r(j)$ of each sensor resistance:

$$r(j) = \frac{R(j) - R_0(j)}{R_0(j)}, \tag{1}$$

where $R_0(j)$ represents the baseline resistance of $j$-th sensor measured in the synthetic air atmosphere.

The measurement system parameters were as follows: a carrier flow 0.2 l/min, a size of the gasoline sample 100 ml, a capacity of the sample chamber 200 ml, a gasoline temperature 25ºC. The sampling rate of sensor resistance was 30 times per minute. The baseline resistance $R_0(j)$ was registered at a stabilized temperature of 25ºC in a synthetic air. Its value was calculated by averaging 36 samples of the measured values within 72 s. A washing interval in the measurement was 10 min. The diagnostic feature vector **x** used in signal pattern recognition was composed of seven relative sensor signals described by (1) and was given in the form $\boldsymbol{x} = [r(1), r(2), ..., r(7)]^T$.

## 3. Data base

The experiments have been performed using pure extracted gasoline, characterized by the following physical and chemical properties: density – 0.665–0.700 g/cm3, final boiling point – 90ºC, relative content of aromatic hydrocarbons – 0.0005% [g/g]. The gasoline has been enriched by different supplements. They included ethanol, *ethyl tert-butyl eter (*ETBE), *methyl tert-butyl eter* (MTBE) and benzene, all of various concentrations in the blend. These components have been applied since they are most often used in petrol industry. Different types of blends, representing the classes under recognition, have been prepared [2]. The first four classes were formed by the extracted gasoline and ethanol concentrations of 5%, 10%, 15% and 20% of the volume. These concentrations have been chosen to reflect the recommended or accepted levels of bio-components in different countries (Brasil – 20%, USA – from 10% to 15%, Poland – 5%). The next four classes were created by adding MTBE and ETBE to the extracted gasoline in the proportion: MTBE (3%) and ETBE (97%). The last four blend families were created by adding benzene as a supplement. Benzene has a high octane number and thanks to this it is an important component added to the gasoline.

Four different blends of 5%, 10%, 15% and 20% concentrations for all mentioned supplements have been created in this way. The detailed description of classes is given in Table 1. Each class is represented by 72 carefully measured samples at a temperature of 25ºC, prepared from two different deliveries. The total number of samples is 864 and their acquisition has been done on the same day. The data analysed here were taken from [2], where only the original, not disturbed measurements, have been considered. They form the nominal set of data.

An additional set of sensor samples has been registered at an increased temperature (by heating the room space) of around 32ºC to observe the influence of temperature changes. These measurements have been done on another day and lasted a few hours. The environmental temperature was slightly changed from 31ºC to 34ºC in a random way. These samples have been normalized using the previously measured baseline resistance of sensors estimated at a basic temperature of 25ºC.

Table 1. The classes of gasoline blend used in the experiments.

| Class | Type of additive |
|---|---|
| 1 | Ethanol additive of 5% volume |
| 2 | Ethanol additive of 10% volume |
| 3 | Ethanol additive of 15% volume |
| 4 | Ethanol additive of 20% volume |
| 5 | Additive of 5% volume (MTBE 3% and ETBE 97%) |
| 6 | Additive of 10% volume (MTBE 3% and ETBE 97%) |
| 7 | Additive of 15% volume (MTBE 3% and ETBE 97%) |
| 8 | Additive of 20% volume (MTBE 3% and ETBE 97%) |
| 9 | Benzene additive of 5% volume |
| 10 | Benzene additive of 10% volume |
| 11 | Benzene additive of 15% volume |
| 12 | Benzene additive of 20% volume |

Figure 1 presents the influence of the environmental temperature on the measured signals from all sensors [19]. The solid line represents the basic results of measurements at a temperature of 25ºC and the dashed line the measurement results made at an increased temperature of around 32ºC.

An important task of the paper is to study the statistical behaviour of the electronic nose system in the existence of the disturbing noise. The zero mean Gaussian distribution white noise of different variance has been assumed to represent the possible measurements made in different environmental conditions. Different noise levels have been used in the experiments. The *signal-to-noise* (SNR) ratio varied from 60 dB to 0 dB. The SNR was defined in a standard way as the logarithm of the ratio of autocorrelation $R_{ss}$ of signal and autocorrelation $R_{nn}$ of noise, $SNR = 10\log\frac{R_{ss}}{R_{nn}}$. The SNR measured totally for all sensor signals has been assumed.

The classification experiments at this point aimed to check the robustness of the nose systems to the possible distortion in the measurement process. The strategy of cross-validation of data has been used in the experiments. The whole data set was split randomly into two equal parts. One part was used in the learning and the second – only in the testing mode. Two types of classification systems have been used: the random forest and the support vector machine. The random splits of data have been repeated 10 times changing the contents of learning and testing subsets. A percentage of testing error is estimated as the mean of the errors committed by the system in all runs.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

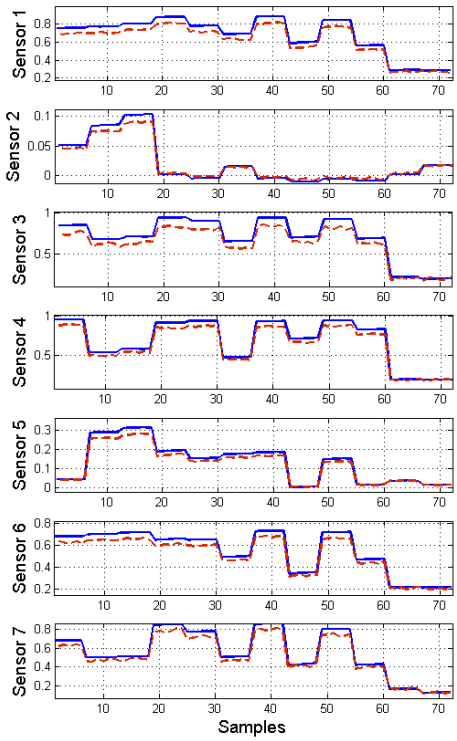



Fig. 1. The graphical effect of changing the environmental temperature for application of the same baseline resistance in both cases. The sensor signals represent normalized (dimensionless) values.

## 4. Self-organizing approach to mining the electronic nose data

### *4.1. Algorithms of clustering*

Clustering of data consists in self-organizing division of $n$ observations in an $N$-dimensional space into $K$ subsets (clusters) represented by their centers $\mathbf{c}_i$, while providing the minimum total distance between the data vectors $\mathbf{x}_j$ and their winning centers $\mathbf{c}_i$ [20]. The process can be either unsupervised (without reference to a class membership of data) or supervised (a known class membership of data). In the case of unsupervised analysis the mathematical problem is described by:

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{\boldsymbol{x}_j \in S_i} \left\| \boldsymbol{x}_j - \boldsymbol{c}_i \right\|^2. \tag{2}$$

The most popular solution is based on the competition technique. In the off-line implementation it is called $K$-means [20], whereas in the on-line mode its modification is known as *Conscience Winner Takes All* (CWTA). In this paper the on-line implementation is used, since it is less susceptible to the so called dead centers. Moreover, the on-line implementation leads usually to a smaller value of quantization error described by (2). In this technique each vector $\mathbf{x}_j$ is associated with its nearest center, which is subject to adaptation according to the relations [20]:

$$\boldsymbol{c}_i(k+1) = \boldsymbol{c}_i(k) + \eta\left[\boldsymbol{x}_j - \boldsymbol{c}_i(k)\right], \tag{3}$$

where $\eta$ is a learning coefficient (usually $\eta < 1$), reduced to zero from iteration to iteration. To avoid the dead centers, the distance between the vector **x** and its winning center takes into account the past activity of the center by multiplying the real distance $d_r(\mathbf{x}_j,\mathbf{c}_i)$ by the number $N_i$ of winnings of $i$-th center in the previous competitions, $d(\mathbf{x}_j,\mathbf{c}_i) \leftarrow N_i d_r(\mathbf{x}_j,\mathbf{c}_i)$ for $i = 1,2,\ldots,K$. At the beginning of the process $N_i = 1$ is assumed for all centers. This modification of distance calculation is applied only in the first cycle of adaptation.

After the segmentation stage the clusters are created, each containing vectors **x** of a high similarity. From a classification point of view the best case is when the data belonging to one class occupy similar locations in the space and are represented by a single cluster.

The quality of unsupervised clustering might be assessed by using different measures [20], from which the most often used are its purity and precision. The purity measures the extent to which a cluster contains objects of the same class. The purity of $i$-th cluster is defined as $p_i = \max_j p_{ij}$, where $p_{ij}$ represents the percentage of representatives of $j$-th class in $i$th cluster. The overall purity of clustering $n$ data points into $K$ clusters is $p = \sum_{i=1}^{K} \frac{n_i}{n} p_i$, where $n_i$ is the population of $i$-th cluster. The class precision is defined as the fraction of a cluster that consists of samples of the specified class. The precision of cluster $i$ with respect to class $j$ is $prec(i,j) = p_{ij}$.

In the case of supervised clusterization the samples belonging to one class are grouped into a separate cluster of the center being the average of data of this class. The quality of supervised clustering is usually assessed on the basis of cluster cohesion (compactness) and separation. The cluster cohesion defines the similarity of data points **x** to their cluster center **c**. It is usually expressed as the averaged distance between the data points and their representative center. On the other hand, the separation between two clusters is characterized by the distance between their centers.

### *4.2. Results of clustering experiments*

The first experiments have been directed to finding relations between the class and the cluster membership of data. The CWTA algorithm was applied. The number of clusters was equal to the number of classes (12 in the experiments). Table 2 presents the percentages of classes belonging to particular clusters for the original data obtained at a nominal temperature of 25ºC. It is seen that all clusters are very good representatives of the classes and their class precision is 100%.

Table 2. The percentage of class membership in particular clusters obtained in CWTA self-organization for the original samples measured at a temperature of 25ºC.

| CLUSTER \ CLASS | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | | | | | | | | | | | |
| 2 | | 100 | | | | | | | | | | |
| 3 | | | 100 | | | | | | | | | |
| 4 | | | | 100 | | | | | | | | |
| 5 | | | | | 100 | | | | | | | |
| 6 | | | | | | 100 | | | | | | |
| 7 | | | | | | | 100 | | | | | |
| 8 | | | | | | | | 100 | | | | |
| 9 | | | | | | | | | 100 | | | |
| 10 | | | | | | | | | | 100 | | |
| 11 | | | | | | | | | | | 100 | |
| 12 | | | | | | | | | | | | 100 |

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

The situation has changed significantly after the measurements performed at an increased temperature of around 32ºC. The measured data have been normalized by using the same value of the previous baseline corresponding to the temperature 25ºC. Table 3 presents the class distribution of such data. Here, we can observe some mixture of classes belonging to the same clusters. Only four clusters (2, 3, 5 and 12) contain the samples of a single class. The class precision of some clusters is low.

This distribution of classes in the presence of temperature changes is an evidence of the influence of the environmental temperature on the signal patterns representing different classes of data.

Table 3. The percentage of class membership in particular clusters obtained in CWTA self-organization after changing the environmental temperature.

| CLUSTER \ CLASS | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 94.4 | | | | | | | | 5.6 | | | |
| 2 | | 100.0 | | | | | | | | | | |
| 3 | | | 100 | | | | | | | | | |
| 4 | | | | 88.9 | | 33.3 | | | | | | |
| 5 | | | | | 100 | | | | | | | |
| 6 | | | | 11.1 | | 66.7 | | | | | | |
| 7 | | | | | | | 80.6 | 11.1 | | 5.6 | | |
| 8 | | | | | | | 11.1 | 80.6 | | 5.6 | | |
| 9 | 5.6 | | | | | | | | 94.4 | | | |
| 10 | | | | | | | | 8.3 | | 88.8 | 8.3 | |
| 11 | | | | | | | 8.3 | | | | 91.7 | |
| 12 | | | | | | | | | | | | 100 |

An important issue in class recognition is the relative distribution of classes, especially distances between class centers. The larger this distance and the smaller standard deviation of data the easier the recognition task. This is especially important in an electronic nose, since the semiconductor sensor signals are vulnerable to noise caused by the change of environmental parameters. Table 4 shows these distances for all class centers (the mean of data belonging to the successive classes) for the samples measured at a nominal temperature. Large changes of these values are observed for different classes.

Similar calculations performed for the data registered at an increased temperature have shown that the centres have not changed their locations in a significant way. The maximum changes of centre locations for the perturbed data did not exceed a relative value of 5%.

Table 4. The distances between class centers for the nominal samples of data.

| **Class** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0.54 | 0.52 | 0.27 | 0.16 | 0.60 | 0.28 | 0.55 | 0.20 | 0.44 | 1.29 | 1.32 |
| 2 | 0.54 | 0 | 0.078 | 0.59 | 0.55 | 0.27 | 0.61 | 0.54 | 0.58 | 0.52 | 0.99 | 1.02 |
| 3 | 0.52 | 0.07 | 0 | 0.55 | 0.52 | 0.32 | 0.57 | 0.57 | 0.54 | 0.54 | 1.05 | 1.08 |
| 4 | 0.27 | 0.59 | 0.55 | 0 | 0.66 | 0.13 | 0.08 | 0.70 | 0.10 | 0.63 | 1.42 | 1.46 |
| 5 | 0.16 | 0.55 | 0.52 | 0.66 | 0 | 0.62 | 0.16 | 0.61 | 0.09 | 0.52 | 1.35 | 1.38 |
| 6 | 0.60 | 0.27 | 0.32 | 0.13 | 0.62 | 0 | 0.71 | 0.36 | 0.67 | 0.42 | 0.79 | 0.82 |
| 7 | 0.28 | 0.61 | 0.57 | 0.08 | 0.16 | 0.71 | 0 | 0.08 | 0.75 | 0.67 | 0.47 | 1.5 |
| 8 | 0.55 | 0.54 | 0.57 | 0.70 | 0.61 | 0.36 | 0.08 | 0 | 0.69 | 0.17 | 0.80 | 0.83 |
| 9 | 0.20 | 0.58 | 0.54 | 0.10 | 0.09 | 0.67 | 0.079 | 0.69 | 0 | 0.60 | 1.42 | 1.45 |
| 10 | 0.44 | 0.52 | 0.54 | 0.63 | 0.52 | 0.42 | 0.67 | 0.17 | 0.60 | 0 | 0.30 | 0.93 |
| 11 | 1.29 | 0.99 | 1.05 | 1.42 | 1.35 | 0.79 | 0.47 | 0.80 | 1.42 | 0.30 | 0 | 0.09 |
| 12 | 1.32 | 1.02 | 1.08 | 1.46 | 1.38 | 0.82 | 1.5 | 0.83 | 1.45 | 0.93 | 0.09 | 0 |

The distances of data samples to their winning center is another important factor that should be taken into account in the analysis. The average distance of samples to their winning center and the standard deviation are significantly dependent on the temperature. Table 5 presents these values for the data registered at nominal and increased temperatures. The third column shows the comparative results for the nominal data distorted by random noise of normal distribution at SNR = 12 dB.

Table 5. The average distances of samples to their centers and standard deviations for the data registered at 25ºC, around 32ºC and the data distorted by random noise of normal distribution at SNR = 12 dB.

| Class | Data registered at temperature 25ºC | Data registered at temperature ~32ºC | Nominal data distorted by random noise of SNR = 12 dB |
|---|---|---|---|
| 1 | 0.005 ± 0.0036 | 0.014 ± 0.0136 | 0.054 ± 0.0321 |
| 2 | 0.005 ± 0.0045 | 0.024 ± 0.0229 | 0.025 ± 0.0230 |
| 3 | 0.003 ± 0.0028 | 0.015 ± 0.0148 | 0.016± 0.0134 |
| 4 | 0.020 ± 0.0143 | 0.024 ± 0.0162 | 0.048 ± 0.0351 |
| 5 | 0.006 ± 0.0046 | 0.018 ± 0.0193 | 0.068 ± 0.0433 |
| 6 | 0.012 ± 0.0074 | 0.015 ± 0.0136 | 0.040 ± 0.0305 |
| 7 | 0.005 ± 0.0044 | 0.026 ± 0.0272 | 0.033 ± 0.0190 |
| 8 | 0.025 ± 0.0168 | 0.028 ± 0.0256 | 0.044 ± 0.0374 |
| 9 | 0.003 ± 0.0023 | 0.030 ± 0.0268 | 0.034 ± 0.0164 |
| 10 | 0.008 ± 0.0057 | 0.023 ± 0.0313 | 0.039 ± 0.0356 |
| 11 | 0.005 ± 0.0035 | 0.017 ± 0.0125 | 0.052 ± 0.0377 |
| 12 | 0.003 ± 0.0022 | 0.019 ± 0.0184 | 0.050 ± 0.0258 |

The results show that the distorted data are characterized by significantly larger average distances from their winning centers (higher dispersion). This increase depends on the amount of noise and – in the case of SNR = 12dB – the largest observed relative increase is almost 17 (for the 12th class of data). It means that in this case some distorted samples are closer to the neighbouring centers than to their own, which is equivalent to a misclassification.

To obtain a graphical presentation of the distribution of classes we have mapped the 7- dimensional data onto two dimensions by applying the *principal component analysis* (PCA) of the measured samples. The PCA [20] is a linear transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$, mapping the $N$- dimensional original vector $\mathbf{x}$ onto a $K$-dimensional output vector $\mathbf{y}$, of $K < N$ (in our case $K = 2$). The transformed vectors $\mathbf{y}$ preserve the most important features of the original information. The PCA matrix $\mathbf{W}$ is composed of the most important eigenvectors of a covariance matrix $\mathbf{R}_{\mathbf{xx}}$ built for the set of input vectors $\mathbf{x}_i$. The 2-dimensional coordinate system in this analysis is created by the first two most important principal components $PC_1 = y_1$ and $PC_2 = y_2$.

Figure 2a shows the results of PCA analysis of the originally (non-disturbed) measured samples. The samples belonging to different classes have been denoted by the letter C and the successive number of class, *i.e.* C1, C2, *etc*. The presented distribution of these mapped samples is an evidence that the points belonging to particular classes of gasoline blends create separated compact clusters. Twelve distinct gasoline clusters, each composed of samples belonging to the same class, are easily recognized. The clusters are well separated from each other and characterized by a relatively small dispersion. This is a very good prognostic for accurate recognition and separation of all classes.

Adding noise to the measured samples introduces fuzziness in the data distribution (Fig. 2b). The clusters representing different classes occupy now a larger space and the samples belonging to the closest clusters interlace each other, making the class recognition problem much more difficult. For example, six classes on the right side of the figure form now a completely mixed

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

environment of data. The higher the level of distortion the more fuzzy character of the cluster distribution is observed. Therefore, applying the clustering in the process of class recognition based on the purity of clusters in a noisy environment is inefficient and leads to a very large degree of misclassification.

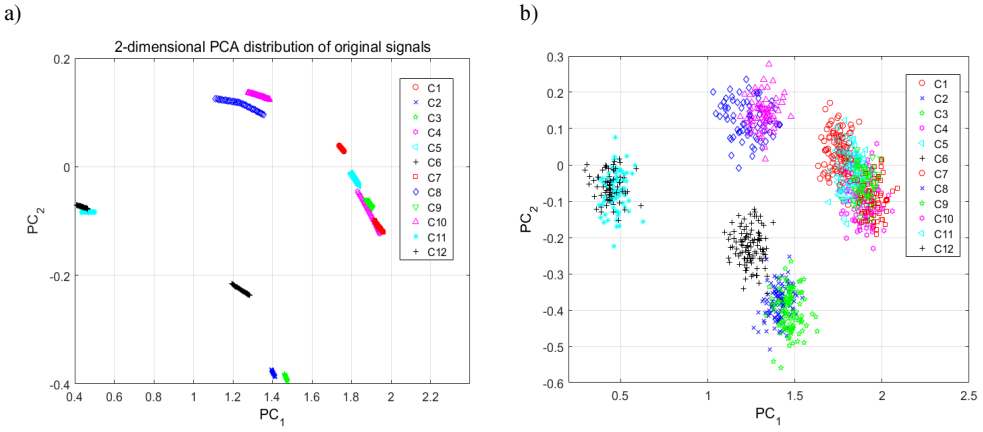



Fig. 2. 2-dimensional PCA plots of the data samples: the original measured data (a); the data corrupted by noise with SNR = 12 dB (b).

## 5. Supervised approach to data mining

### *5.1. Supervised classifiers*

The supervised approach includes application of two types of classifiers: the *random forest* of decision tree (RF) and the *support vector machine* (SVM).

The random forest proposed by Breiman [11] is an ensemble of many multivariate decision trees. It constructs decision trees in the training time and outputs a class, that is the mode of classes pointed by individual trees (majority voting). The learning data (often 2/3 of the whole data set) are selected randomly for each tree.

A small group of input variables to split is selected at random in each node of a decision tree. The group size $m$ is fixed. The linear combinations of $m$ randomly selected variables in each node are generated, and then a search is made over them for the best split. The predicted variables that provide the best split according to a predefined objective function are used to do a binary split on that node. In the next nodes, other $m$ variables are chosen at random from all predicted variables and previous operations are repeated. After generating a large number of trees they vote for the most popular class.

It was proved [11] that random split selection of data increases the recognizing system immunity to noise contained in the measurement data. This is a very important property of an electronic nose, since the sensor signals may experience some unpredictable variations.

The application of RF enables to assess the significance of individual sensors for the performance quality of an electronic nose. The impact of a particular sensor signal is estimated by taking into account its influence on the classification results, in particular, how inclusion of this signal is important for getting a higher accuracy of class recognition [11].

Generally, the importance of input attribute in RF is measured by an increase of prediction error for the validation data if the values of this attribute are permuted among the testing data. The out-of-bag prediction error is computed on this perturbed data set and compared with the error before perturbation. The higher this increase, the more important is the input attribute.

This measure is estimated for every tree, then averaged over the ensemble and divided by the standard deviation over the entire ensemble [11, 19]. In this way the input attributes (sensor signals) are ordered according to their statistical impact on the classification accuracy.

The *Support Vector Machine* (SVM) is a feedforward network of one hidden layer (the kernel function layer) known for its good generalization ability [12]. In the learning phase it constructs a hyperplane in a high-dimensional space, separating the learning vectors into two classes of the destination values either $d_i = 1$ (one class) or $d_i = -1$ (the opposite class), with the maximal separation margin (the largest distance of the nearest training data points of the opposite classes). The SVM model represents the original data as points mapped in the space in such a way that examples of different categories are separated by clear gaps that are as wide as possible. The width of separation margin formed in the learning stage depends on a regularization constant $C$, which should be properly adjusted by the user. Thanks to such a learning strategy the network is resistant to noise contaminating the input data.

A great advantage of SVM is unique formulation of the learning problem leading to the quadratic programming with linear constraints, which is easy to solve. The SVM of Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|)$, treated as the most universal one, has been used in this application. The hyper-parameters $\gamma$ of the Gaussian function and the regularization constant $C$ have been adjusted by repeating the learning experiments for the set of their predefined values and choosing the best one in the validation data sets. The optimal values of these parameters found in the preliminary experiments were as follows: $\gamma = 1$ and $C = 1000$. They have been found by applying a trial-and-error process using the validation data (a third of the learning data volume). To deal with the problem of many classes we have applied a one-against-one strategy [12]. In this method many 2-class SVM classifiers are used for all combinations of two classes. The final classification decision is based on the majority voting principle.

### *5.2. Supervised classification results*

The recognition ability of the classes of blends was checked by applying two systems of supervised classification: the support vector machine and the random forest of decision trees. The SVM of Gaussian kernel and hyper parameters $\gamma = 1$ and $C = 1000$ were applied in all 10 validation runs. On the other hand, the RF of 50 trees was constructed using 3 input variables selected at random in each node to split. These meta-parameter values of RF have been adjusted after some introductory experiments using the trial-and-error approach, in which different values of trees and node variables have been tried. The choice providing the best results on the validation data (a third of the learning data volume) has been accepted in further experiments. The RF experiments with random selection of learning and testing data have been also repeated 10 times and their results averaged.

Application of both classifiers to the recognition of original samples of blends registered at a temperature of 25°C has resulted in 100% accuracy of class recognition. These excellent results are in accordance with the class-cluster distribution presented in Table 2 and also with the PCA results of the original data presented in Fig. 2a.

However, introducing noise to the testing samples while training the classifiers on an undisturbed data set, has resulted in decreasing this accuracy. This reduction was dependent on the actual SNR value, the type of applied classification system and the number of applied sensors. Generally, the higher the noise, the larger degree of misclassification.

An important issue is the impact of individual sensors on the class recognition results. In solving this problem the random forest ability was used. A measure of the sensor importance in this method is defined as a relative increase of error after perturbation of its value compared with the error before this perturbation The more important sensor corresponds to a larger

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

relative increase of this error. Fig. 3 presents the result of application of random forest to the measure of class discrimination ability of successive sensors. The results show that all sensors contribute positively to the classification results and their influence is at a similar, although not equal, level.

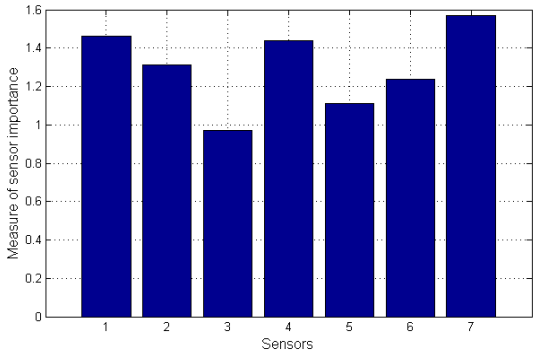



Fig. 3. The importance of sensors in class recognition according to the random forest measure.

The next experiments have been performed with a gradually reduced number of sensors. In the sensor elimination process the embedded property of random forest to assessing the importance of individual sensor signals in the classification process has been used. The results of this estimation are presented in Fig. 3. The least important in this phase of classification was the third sensor. Therefore, in the next runs of calculations this sensor signal was eliminated from the input vector as the first one. Next, the same procedure of estimation of importance of the remaining six sensors was repeated which to elimination of the next least important sensor (the importance of sensors is changing from run to run and depends on a set of remaining sensors).

The statistical results concerning the recognition error of the blend classes at different levels of noise and different quantities of applied sensors are presented in Table 6. The classifiers were trained on the undisturbed nominal data subsets and tested on the other subsets disturbed by artificial noise of different levels. The results refer to application of the SVM and RF classification systems to three different matrices of sensors, *i.e*. containing 7, 6 and 5 sensors. The table presents the mean values and *standard deviations* (std) of class recognition obtained in 10 runs of the classification process. Each run was associated with random noise of a specified SNR value.

There is a visible correlation between the number of applied sensors and robustness of the system to the noise level. The elimination of sensors leads to an increase of the recognition error. A type of classifier is also very important. The results show that SVM is much more resistant to noise than the random forest. This is well seen for high levels of noise. For example, at SNR = 20 dB and application of 7 sensors the mean error of class recognition for SVM was equal to 5.12% and for RF it increased to 19.87%. In the case of six sensors the respective values were 5.80% (SVM) and 21.14% (RF). An advantage of SVM over RF follows from the margin of separation built in the learning process of SVM, which is not the case for RF.

Figure 4 shows a plot of the mean values of relative recognition error versus SNR values for both classifiers and different numbers of sensors applied. These results have been obtained in 10 runs of the classification procedure. They confirm the superiority of SVM performance in the presence of noise.

Table 6. The mean errors of recognition of gasoline blends in application of RF and SVM for different levels of noise.

| SNR [dB] | Mean error and std for 7 sensors [%] | | Mean error and std for 6 sensors [%] | | Mean error and std for 5 sensors [%] | |
|---|---|---|---|---|---|---|
| | SVM | RF | SVM | RF | SVM | RF |
| 60 | 0 | 0 | 0 | 0.20 ± 0.12 | 0 | 0.30 ± 0.18 |
| 45 | 0 | 0.26 ± 0.19 | 0 | 0.32 ± 0.30 | 0 | 0.82 ± 0.25 |
| 52 | 0 | 0.80 ± 0.51 | 0 | 0.90 ± 0.53 | 0 | 1.01 ± 0.68 |
| 32 | 0 | 6.12 ± 1.62 | 0 | 8.71 ± 2.50 | 0 | 9.95 ± 1.23 |
| 25 | 0.37 ± 0.32 | 13.23 ± 3.4 | 0.39 ± 0.34 | 17.5 ± 1.59 | 0.56 ± 0.42 | 21.45 ± 3.63 |
| 20 | 5.12 ± 1.34 | 19.87 ± 3.26 | 5.80 ± 1.68 | 21.14 ± 2.84 | 6.9 ± 0.98 | 26.34 ± 3.82 |
| 15 | 10.87 ± 2.34 | 38.47 ± 3.78 | 19.68 ± 2.08 | 40.49 ± 3.92 | 22.78 ± 2.07 | 43.87 ± 4.03 |
| 10 | 27.21 ± 3.24 | 42.09 ±2.73 | 28.43 ± 4.02 | 44.34 ± 2.13 | 31.67 ± 4.62 | 47.58 ± 2.24 |
| 7.5 | 35.35 ± 2.41 | 53.24 ± 3.96 | 36.78 ± 2.64 | 55.47 ± 4.81 | 39.68 ± 2.34 | 60.56 ± 3.15 |
| 6 | 41.24 ± 3.24 | 62.89 ± 3.68 | 44.21 ± 3.91 | 65.39 ± 3.58 | 45.23 ± 3.17 | 68.12 ± 4.25 |
| 0 | 60.34 ± 2.95 | 80.25 ± 2.73 | 62.80 ± 1.98 | 83.73 ± 1.99 | 64.52 ± 3.02 | 85.89 ± 3.27 |

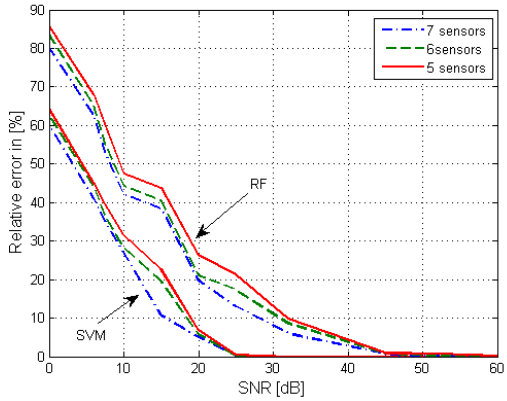



Fig. 4. The relative error of class recognition by RF and SVM for different levels of noise corrupting the measured data and different numbers of sensors.

## 6. Wavelet de-noising sensor signals

The nose measurements made in various environmental conditions produce non-deterministic sensor signals in different frequency domains with respect to the basic measurement, at which the baseline was estimated. Therefore, a proper transformation of sensor signals from the time to the frequency domain and a careful removal of noisy components can filter the sensor signals. The procedure presented in the paper was implemented by using the discrete wavelet transform.

### *6.1. Principle of wavelet de-noising*

The *discrete wavelet transform* (DWT) is a linear transformation with a special property of simultaneous localization in time and frequency. It decomposes a given discrete signal series into a set of specially defined basic functions of different frequencies, shifted mutually and called wavelets [21, 22].

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

The aim of DWT is to decompose an analyzed signal $x(t)$ into a finite summation of wavelets of different scales (levels) and shifts according to the expansion:

$$x(t) = \sum_j \sum_k c_{jk} \psi(2^j t - k), \tag{4}$$

where $c_{jk}$ is a new set of coefficients and $\psi(2^j t - k)$ is the wavelet of $j$-th level (scale) shifted by $k$ samples. The set of wavelets of different scales and shifts can be generated from a single prototype mother wavelet, by dilations and shifts. In practice, most often used are the orthogonal or bi-orthogonal wavelet functions, forming an orthogonal or bi-orthogonal base [19, 21].

Denote a discrete form of the original signal vector by $\mathbf{x}$ and by $A_j\mathbf{x}$ an operation that computes the approximation of $\mathbf{x}$ with a resolution $2^j$. Let $D_j\mathbf{x}$ denotes the detailed signal, $D_j\mathbf{x} = A_{j+1}\mathbf{x} - A_j\mathbf{x}$, defined as the difference of signal approximations for two neighboring resolutions. Both operations $A_j\mathbf{x}$ and $D_j\mathbf{x}$ can be interpreted as the convolution of the signal of previous resolution and the finite impulse response of the quadrature mirror filters: the high-pass one ($\widetilde{G}$) with coefficients $\tilde{g}$ and the low-pass one ($\widetilde{H}$) with coefficients $\tilde{h}$:

$$A_j \boldsymbol{x} = \sum_{k=-\infty}^{\infty} \tilde{h}(2n-k) A_{j+1} \boldsymbol{x}(2n), \tag{5}$$

$$D_j \boldsymbol{x} = \sum_{k=-\infty}^{\infty} \tilde{g}(2n-k) A_{j+1} \boldsymbol{x}(2n). \tag{6}$$

These two operations performed at different levels, from $j = 1$ to $J$, deliver the decomposition coefficients for different scales and resolutions of the original signal $\mathbf{x}$. The most often used discrete wavelet analysis uses the Mallat pyramid algorithm [22].

The result of such transformation is a set of coefficients representing the detailed signals $D_j\mathbf{x}$ at different levels $j$ ($j$ = 1, 2,.., $J$) and the residual signal $A_J\mathbf{x}$ at the level $J$. All of them are of different resolution, characteristic for the applied level. The $D_j\mathbf{x}$ can be interpreted as the high frequency details, distinguishing the approximations of the signal for two neighboring levels of resolution. The signal $A_J\mathbf{x}$ represents a coarse approximation of the vector $\mathbf{x}$.

Transformation of the detailed signals $D_j\mathbf{x}$ ($j$ = 1, 2,..., $J$) and the coarse approximation signals $A_J\mathbf{x}$ into the original resolution is possible using special filters $G$ and $H$ associated with the analysis of filters $\widetilde{G}$ and $\widetilde{H}$ by the quadrature and reflection relationships [21, 22]. This is done by the reverse Mallat pyramid algorithm. The original signal $\mathbf{x}(n)$ at each time instant $n$ is reconstructed by simply adding appropriate wavelet coefficients and the coarse approximation, both transformed to the same original resolution. At $J$-th level of DWT we have:

$$x(n) = D_1(n) + D_2(n) + \ldots + D_J(n) + A_J(n). \tag{7}$$

Figure 5 presents the results of 4-level DWT of the sensor data in measurement of petrol with bio-additives after adding noise [19].

The Haar wavelets have been applied in the decomposition. The first four levels of wavelet coefficients represent the detailed coefficients from $D_1$ to $D_4$, whereas the next one – denoted by $A_4$ – the coarse approximation at the 4th level. All are presented in the original resolution. We observe a substantial difference of variability of signals at different levels. The first level detail coefficients $D_1$ represent the highest variability of signal, which is usually associated with the high frequency noise.

Application of the wavelet transformation in sensor technology is not new. It was used *e.g.* for extraction of features in porous silicon chemical sensors [23]. Our idea is to apply this transformation to reduction of noise. This is done by cutting the lowest value detailed coefficients of wavelet decomposition and reconstruct the sensor signals deprived of them. The cut coefficients are treated as the noise components. Thanks to this we obtain reduction of noise contaminating the measured signals.

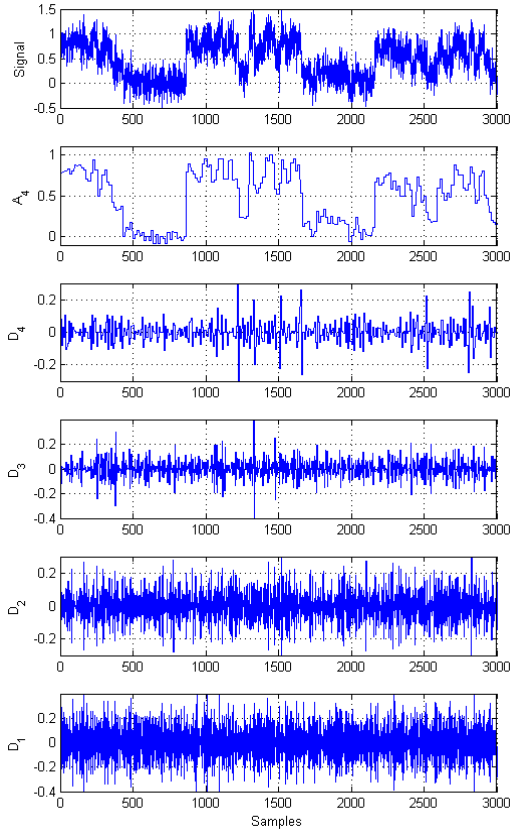



Fig. 5. DWT of the measured sensor signals. $D_1$ to $D_4$ represent the detailed coefficients and $A_4$ the coarse approximation of signal at the 4th level. The signals are represented by the normalized (dimensionless) values.

### *6.2. The results of de-noising using DWT*

The next experiments have been performed using DWT for de-noising the sensor signals artificially distorted by the white noise with a normal distribution and a different variance. The aim is to reduce the noise in the data and in this way to increase the probability of proper pattern recognition. The de-noising process is performed by decomposing the noisy sensor signals into a few decomposition levels and then reconstructing them by eliminating the least important details. The four-level decomposition using Haar wavelets and soft thresholding of fixed values in each detail coefficient have been found to be the best. The results of such de-noising are presented in Fig. 6.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

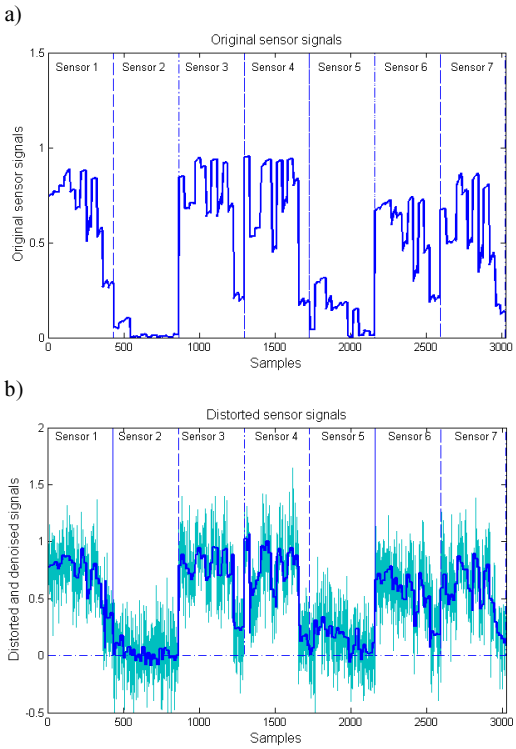



Fig. 6. An illustration of the de-noising process of sensor signals: the original samples of undistorted sensor signals (a); the de-noised signal (blue color) with the noisy signal at SNR = 0 dB at the background (green color) (b). The signals are represented as the normalized (dimensionless) values.

The upper Fig. 6a presents the original samples of undistorted sensor signals and the bottom one (Fig. 7b) the de-noised samples with the distorted signals at SNR = 0 dB at the background. A significant decrease of noise is visible. Fig. 7 illustrates the effect of signal de-noising by presenting the relative difference between the original sensor signals $x(n)$ and the distorted signals $x_{noise}(n)$ and between the original signals and the signals after de-noising $x_{denoise}(n)$. This difference is expressed in the form of relative coefficients $\alpha_{noise}$ and $\alpha_{denoise}$ :

$$\alpha_{noise} = \frac{\|x(n) - x_{noise}(n)\|}{\|x(n)\|}, \tag{8}$$

$$\alpha_{denoise} = \frac{\|x(n) - x_{denoise}(n)\|}{\|x(n)\|}. \tag{9}$$

The effect of de-noising is visible, especially at high values of noise (SNR close to zero). The reduction ratio of noise contents in the signal in such a case well exceeds 2.

This effect is also very well seen in the 2-dimensional coordinate system formed by two most important PCA components of the sensor signals. This is illustrated in Fig. 8.

Figure 8a presents the distribution of noisy samples of the sensor signals at SNR = 0 dB and Fig. 8b their distribution after the de-noising process. In the first case (Fig. 8a) the samples belonging to different classes are completely mixed, while after de-noising (Fig. 8b) the representatives of individual classes are grouped together and the classes are reasonably well separated from each other.

The de-noised sensor signals have been used as the input attributes to the SVM classifier in the process of class recognition. The classifier was trained on the original (undistorted) part

of data and then tested on a separate part of distorted data before and after de-noising. The average results of 10 runs in such organized classification process at different noise levels are presented in Table 7.

The table presents the mean percentage errors and their standard deviations obtained in 10 runs of the classification process for the noisy data before and after their de-noising. The results are given for 7 sensors and after reducing their numbers to 6 and 5, preserving the most important sensors. A significant reduction of classification errors can be observed for each level of noise.

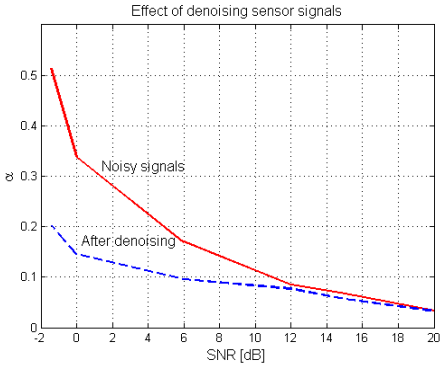



Fig. 7. The influence of the de-noising process on the reduction of noise contents in the signals.

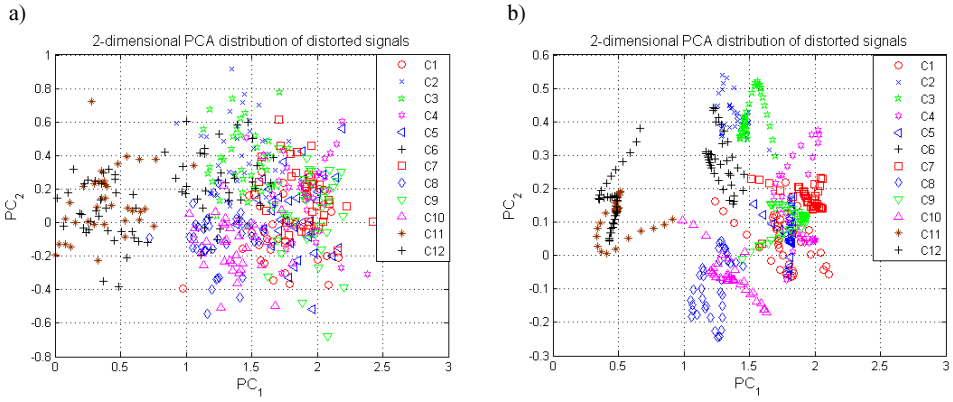



Fig. 8. The distribution of sensor signal samples of 12 classes mapped onto the 2-dimensional coordinate system formed by two most important principal components $PC_1$ and $PC_2$: signals distorted by noise of SNR = 0 dB (a); signals after the DWT de-noising process (b).

Table 7. The SVM classification results of de-noising.

| SNR [dB] | Mean error and std for 7 sensors [%] | | Mean error and std for 6 sensors [%] | | Mean error and std for 5 sensors [%] | |
|---|---|---|---|---|---|---|
| | noisy signals | after de-noising | noisy signals | after de-noising | noisy signals | after de-noising |
| 20 | 5.12 ± 1.34 | 2.7 ± 0.63 | 5.80 ± 1.68 | 3.5 ± 0.71 | 6.9 ± 1.17 | 4.7 ± 0.98 |
| 15 | 10.87 ± 2.34 | 4.4 ± 0.91 | 19.68 ± 2.08 | 6.3 ± 1.03 | 22.78 ± 2.07 | 10.9 ± 1.34 |
| 12 | 27.21 ± 3.24 | 12.3 ± 1.67 | 28.43 ± 4.02 | 12.9 ± 1.89 | 31.67 ± 4.62 | 14.7 ± 2.23 |
| 6 | 41.24 ± 3.24 | 18.1 ± 2.54 | 44.21 ± 3.91 | 19.2 ± 2.79 | 45.23 ± 3.17 | 26.5 ± 2.96 |
| 0 | 60.34 ± 2.95 | 35.1 ± 1.71 | 62.80 ± 1.98 | 35.9 ± 1.87 | 64.52 ± 3.02 | 36.3 ± 2.03 |

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 27–44.

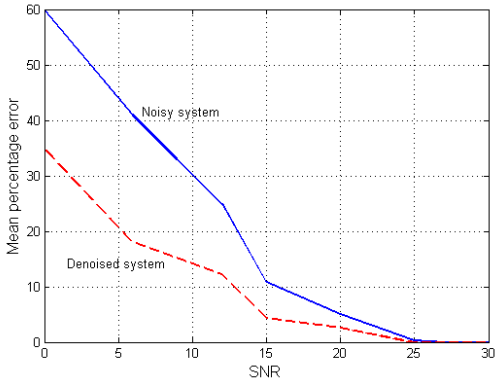



Fig. 9. The dependence of the mean recognition error on the SNR value for 7 noisy sensor signals before and after their de-noising.

Figure 9 shows this difference for 7 sensors in a graphical way. It presents the dependence of the mean percentage value of class recognition error on signal de-noising obtained for all 7 sensors obtained for the noisy sensor signals .

## 7. Conclusions

The paper presents the analysis of data distorted by noise, simulating a noisy environment of electronic nose measurement of the gasoline blends. The array of semiconductor sensors, forming the heart of an electronic nose, corresponds with a signal pattern characteristic for each gasoline blend type. The pattern recognition system working in the classification mode processes these signals and associates them with an appropriate class. The problem is that each nose measurement includes some distortion resulting from changing environment conditions and fluctuation of sensor signals. The examinations aimed in two directions: the unsupervised analysis based only on the measured sensor signals, and the supervised analysis, where the class represented by a signal pattern is known in the learning phase. It was proved that the noise corrupting the measurement has a significant influence on the locations of clusters, and also increases their dispersion, making the pattern recognition based on distances not efficient.

Therefore, the research was directed to finding the most efficient supervised classification systems. Basing on the actual experience in this field two most efficient classifiers have been chosen: the RF and SVM ones. The main advantage of RF is the fact that it never over fits. Injecting a right kind of randomness in each step of signal processing at a large number of grown trees improves the generalization ability of the system and makes RF a good solution to the classification problems. On the other hand, the SVM network is resistant to noise because of the maximized separation margin between two recognized classes, formed in the learning phase. Such a property makes this tool suitable for working in noisy environments.

Both classification systems have been tested on samples of the gasoline blends formed on the basis of a few bio-based additives injected in different concentrations. In the case of original (noiseless) data, 100% accuracy of class recognition irrespectively of the applied classifiers has been achieved. However, adding noise to the measured samples has decreased the accuracy. In this case SVM has shown its superiority over RF, because of its large insensitivity to noise due to the separation margins between different classes introduced in the learning phase.

The results of class recognition in the presence of noise have confirmed that an electronic nose built on the basis of SVM is a better solution to the pattern recognition of gasoline blends.

The last direction of research was to examine the possibility of de-noising sensor signals. The DWT has been applied in this process. By cutting the small (insignificant) detail

coefficients a great reduction of noise has been achieved. Such de-noised sensor signals applied as the input attributes to the classifiers have increased the accuracy of pattern recognition and the efficiency of final classification system.

## References

[1] McCarrick, C.W., Ohmer, D.T., Gillil L.A., Edwards, P.A. (1996). Fuel identification by neural network analysis of the response of vapour-sensitive sensor arrays. *Analytical Chemistry*, 68, 4264–4269.

[2] Brudzewski, K., Osowski, S., *et al*.(2006). Classification of gasoline with supplement of bio-products by means of an electronic nose SVM neural network. *Sensors and Actuators B*, 113, 135–141.

[3] Di Natale, C., Martinelli, E., D'amico, A. (2005). Pre-processing and pattern recognition methods for artificial olfaction systems: a review. *Metrol. Meas. Syst*., 12(1), 3–26.

[4] Bielecki, Z., Janucki, J., *et al*. (2012). Sensors and systems for the detection of explosive devices – an overview. *Metrol. Meas. Syst.*, 19(1), 3–28.

[5] Boeker, P. (2014). On 'Electronic Nose' methodology. *Sensors and Actuators B*, 204, 2–17.

[6] Jha, S.K., Yadava, R.D. (2011). Denoising by singular value decomposition its application to electronic nose data processing. *IEEE Sensors Journal*, 11, 1, 35–44.

[7] Hassanpour, H. (2008). A time-frequency approach for noise reduction. *Digital Signal Processing*, 18, 728–738.

[8] Fonollosa, J., Fernández, L., *et al*. (2016). Calibration transfer and drift counteraction in chemical sensor arrays using direct standardization. *Sensors and Actuators B*, 236, 1044–1053.

[9] Zuppa, M., Distante, C., Siciliano, P., Persaud, K.C. (2004). Drift counteraction with multiple self-organizing maps for an electronic nose. *Sensors and Actuators B*, 98, 305–317.

[10] Kalinowski, P., Jasiński, G., Jasiński, P. (2014). Stabilność odpowiedzi półprzewodnikowych czujników gazu w zmiennych warunkach środowiskowych: badania długoterminowe oraz korekcja dryftu. *Elektronika: konstrukcje, technologie, zastosowania*, 55(9), 119–121.

[11] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

[12] Schölkopf, B., Smola, A. (2002). *Learning with kernels*. Cambridge MA: MIT Press.

[13] Wiziack, N.K.L., Catini, A., Santonico, M., D'Amico, A., Paolesse, R., Paterno, L.G., Fonseca, F.J., Di Natale, C. (2009). A sensor array based on mass capacitance transducers for the detection of adulterated gasolines. *Sensors and Actuators B*, 140, 508–513.

[14] Osowski, S., Tran Hoai, L., Brudzewski, K. (2004). Neuro-fuzzy TSK network for calibration of semiconductor sensor array for gas measurements. *IEEE Trans. on Measurements Instrumentation*, 53, 330–637.

[15] Guney, S., Atasoy, A. (2012). Multiclass classification of n-butanol concentrations with k-nearest neighbor algorithm support vector machine in an electronic nose. *Sensors Actuators B*, 166–167, 721–725.

[16] Botre, B.A., Gharpure, D.C., Shaligram, A.D. (2010). Embedded electronic nose supporting software tool for its parameter optimization. *Sensors and Actuators B*, 146, 453–459.

[17] Pardo, M., Sberveglieri, G. (2005). Classification of electronic nose data with support vector machines. *Sensors and Actuators B*, 107, 730–737.

[18] Liu, M., Wang, M., Wang, J., Li, D. (2013). Comparison of random forest, support vector machine back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage Chinese vinegar. *Sensors and Actuators B*, 177, 970–980.

[19] Matlab user manual (2014). Natick, USA: MathWorks.

[20] Tan, P.N., Steinbach, M., Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson Education Inc.

[21] Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM, Philadelphia.

[22] Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–692.

[23] Murguía, J.S., Vergara, A., Vargas-Olmos, C., Wong, T.J., Fonollosa, J., Huerta, R. (2013). Two-dimensional wavelet transform feature extraction for porous silicon chemical sensors. *Analytica Chimica Acta*, 785, 1–15.

www.czasopisma.pan.pl PAN www.journals.pan.pl
POLSKA AKADEMIA NAUK

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 45–56.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# A NEW APPROACH TO MEASUREMENT OF FREQUENCY SHIFTS USING THE PRINCIPLE OF RATIONAL APPROXIMATIONS

**Fabian N. Murrieta-Rico[1)], Vitalii Petranovskii[2)], Oleg Y. Sergiyenko[3)], Daniel Hernandez-Balbuena[4)], Lars Lindner[3)]**

*1) Posgrado en Física de Materiales, Centro de Investigación Científica y Educación Superior de Ensenada, Carretera Ensenada-Tijuana 3918, Zona Playitas, 22860 Ensenada, B. C., México (✉ fmurriet@cicese.edu.mx, +52 01 646 175 0500)*

*2) Universidad Nacional Autónoma de México, Centro de Nanociencias y Nanotecnología, Carretera Tijuana-Ensenada Km. 107, Pedregal Playitas, 22860 Ensenada, B. C., México (vitalii@cnyn.unam.mx)*

*3) Universidad Autónoma de Baja California, Instituto de Ingeniería, Calle de la Normal S/N y Blvd. Benito Juárez Col. Insurgentes Este Mexicali B. C., México (srgnk@uabc.edu.mx, lindner.lars@uabc.edu.mx)*

*4) Universidad Autónoma de Baja California, Facultad de Ingeniería, Calle de la Normal S/N y Blvd. Benito Juárez Col. Insurgentes Este Mexicali B. C., México (dhernan@uabc.edu.mx)*

**Abstract**

When a frequency domain sensor is under the effect of an input stimulus, there is a frequency shift at its output. One of the most important advantages of such sensors is their converting a physical input parameter into time variations. In consequence, changes of an input stimulus can be quantified very precisely, provided that a proper frequency counter/meter is used. Unfortunately, it is well known in the time-frequency metrology that if a higher accuracy in measurements is needed, a longer time for measuring is required. The principle of rational approximations is a method to measure a signal frequency. One of its main properties is that the time required for measuring decreases when the order of an unknown frequency increases. In particular, this work shows a new measurement technique, which is devoted to measuring the frequency shifts that occur in frequency domain sensors. The presented research result is a modification of the principle of rational approximations. In this work a mathematical analysis is presented, and the theory of this new measurement method is analysed in detail. As a result, a new formalism for frequency measurement is proposed, which improves resolution and reduces the measurement time.

Keywords: frequency measurement, rational approximations, sensors.



## 1. Introduction

*Frequency Domain Sensors* (FDS) are input transducers, which change their frequency output when they are under stimulation of a physical variable. This change is known as a frequency shift. FDS are also known as frequency output sensors. In the last years, the technology standards require performing highly accurate measurements in a short time. Specifically, FDS applied to detection of chemicals and measurement of concentration are being actively researched [1–6]. It is well-known that a gas-sensing device for toxic gases requires a high sensor sensitivity and a quick response [7]. After a careful review of the literature, where we focused on applications of the frequency measurement devices, we have learned that sensors employed in detection of chemicals are very sensitive (1 Hz corresponds to 4.3 ng/cm$^2$ [8]), but they require a long time for measuring. Typically, hundreds of seconds.

The principle of rational approximations is a method of frequency measurement, where the time required for measurement depends on the standard and the measurand. In other words, the higher the unknown frequency value, the shorter the time required for measurement. One of the main advantages of this method (compared with other well-known techniques [9–11])

Article history: received on May 18, 2016; accepted on Aug. 28, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0007.

is a high speed of measurement without diminishing its accuracy. Also, it is insensitive to most common sources of uncertainty in time-frequency measurement systems [12].

In the preliminary works, our research group has explored the application of the principle of rational approximations for measurement of frequency shifts [13, 14]. So far our proposal requires knowing *a priori* the value of a sensor's output before occurring a frequency shift. For this reason and due to defects of fabrication, there is required measurement of a sensor's output before the frequency shift occurrence. In consequence, to measure a frequency shift at the sensor output, at least two measurements – in two different and separated time intervals – are required.

In this work, we show a modification of the rational approximations principle, and its measurement theory is expanded. As a result, after applying the principle of rational approximations, a frequency shift can be measured at the moment of its occurrence.

## 2. Fast frequency measurements using the principle of rational approximations

The principle of rational approximations is a method of frequency measurement. It is based on the number theory, in particular – on a property of rational numbers: the mediant fractions [15]. The principle of rational approximations has many outstanding properties, including: its invariance to jitter, the accuracy limited only by the stability of a reference signal, and the measurement time decrease with the increase of the measurand value increases.

The frequency measurement using the principle of rational approximations is performed by comparing two signals: a reference one ($S_0$) whose frequency value is known ($f_0$) and a signal to measure ($S_x$) with an unknown frequency ($f_x$). Both signals have corresponding periods $T_0$, $T_x$. After a process of signal conditioning [13] the pulses in both signals must have the same pulse width ($\tau$), which must be $\tau \leq T_0/2$ [16, 17].

When the signals are compared, a pulse train of coincident pulses ($S_x$ & $S_0$) is generated. The frequency measurement process starts at the moment of the first coincidence of pulses. Also, simultaneously starts counting the pulses in $S_x$, $S_0$ . The value of $\tau$ defines the duration of coincidences; this effect is analysed and reported in [16]. The numbers of pulses in $S_x$, $S_0$ are denoted by $P_n/Q_n$, where $n$ is the number of coincidence (Fig. 1).

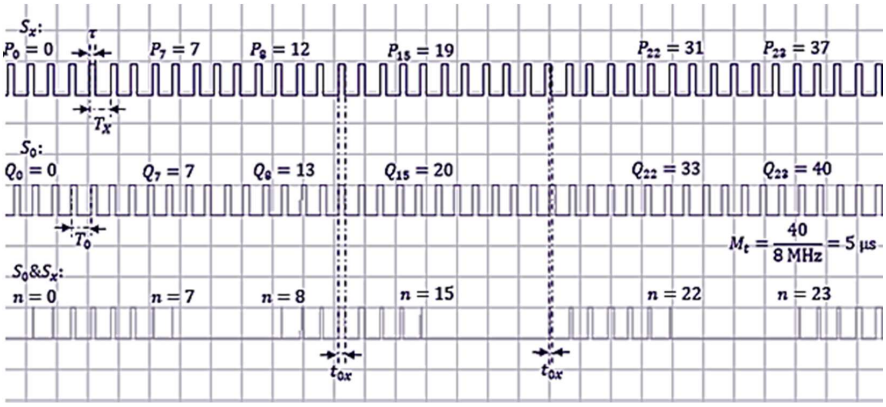



Fig. 1. Comparison of signals during the frequency measurement process, where $f_x = 7.4$ MHz, $f_0 = 8$ MHz, and $\tau \approx 40$ ns.

In each coincidence, a fraction is formed ($P_n/Q_n$), and an approximation to the measurand is obtained. The unknown frequency value is given by:

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 45–56.

$$f_x = \frac{P_n}{Q_n} f_0 \tag{1}$$

or by the sum of all numerators and denominators that form the mediant fraction $m$:

$$f_x = f_0 \frac{\sum_m P_n}{\sum_m Q_n} . \tag{2}$$

The concept of mediant fraction is a well-known property in the number theory, and its application to frequency measurement is shown in (1) of [15]. The measurement time ($M_t$) is given by:

$$M_t = Q_n T_0 = \frac{Q_n}{f_0} . \tag{3}$$

For the frequency measurement process illustrated in Fig. 1, the best approximation to the measurand is obtained in the 23rd fraction, where the second perfect coincidence exists. As the concept of mediant states, all fractions between perfect coincidences are approximations to the measurand. For a continuous pulse counting process, the relative error ($\beta$) in the frequency measurement process is defined as:

$$\beta = \frac{\left| f_x - \frac{P_n}{Q_n} f_0 \right|}{f_x} . \tag{4}$$

From the statements exposed in this section, we have reviewed the basics of the principle of rational approximations. In short, only by counting the pulses after the first coincidence, and in each posterior coincidence, from (2), the value of $f_x$ is calculated. In the next section, the measurement process parameters will be examined and a modification of the principle of rational approximations will be proposed.

## 3. Simultaneous measurement of two frequencies using the principle of rational approximations

Determining an FDS frequency requires at least two measurements. In order to know how the sensor's output changes after a stimulus, measurement of the initial frequency is needed. This process requires two measurements, where the sensor output is measured twice – before and after the stimulus – in different time intervals. In this section, we introduce an original approach, not previously published, to measuring frequency shifts occurring in a given FDS under a stimulus.

When the FDS output has changed, there is a frequency shift of its frequency value. The last can be expressed as:

$$\Delta f = f_s - f_p, \tag{5}$$

where $f_s$ denotes the initial frequency of the sensor; and $f_p$ is the posterior frequency value. These signals have the corresponding periods $T_s$ and $T_p$.

Two sensors of the same kind can be measured simultaneously, and the difference among them can be used to calculate the corresponding frequency shift. If the measured values of sensors are $f_{xs}$ and $f_{xp}$, they can be calculated from measurements using the principle of rational approximations:

$$f_{xs} = \frac{P_{ns}}{Q_{ns}} f_0, \tag{6}$$

$$f_{xp} = \frac{P_{np}}{Q_{np}} f_0. \tag{7}$$

The signal coincidence process of $S_{xs}$ & $S_0$ and $S_{xp}$ & $S_0$ is illustrated in Fig. 2. In this case $f_{xs} = 4.7$ MHz , $f_{xp} = 4.6$ MHz and $f_0 = 8$ MHz. These values are chosen because $f_{xs} = 4.7$ MHz is a common value in *quartz crystal microbalances* (QCMs) [18, 19]. According to the Sauerbrey equation [20], when the frequency is shifted, the frequency value in the QCM output decreases; this is the reason of choosing $f_{xp} = 4.6$ MHz ($f_{xp} < f_{xs}$). Also, until now, the principle of rational approximations states that if the reference frequency is closer to the measurand, the best approximations are obtained in a shorter time [16]. This is why the reference frequency is chosen to be $f_0 = 8$ MHz – that is also a common value in quartz crystals used as the time reference.

An important remark about the graphs of Figs. 2, 3 is required. Simultaneous observing all required parameters of six signals is complicated, because the oscilloscopes (even in simulations) have maximum four channels. In our analysis it is important to evaluate the behaviour of signals from the beginning of measurement process. This enables to evaluate how factors, like the phase or amplitude, affect the frequency measurement process. For the analysed cases, the input signals are digitalized, and the amplitude has just discrete values, but for very narrow pulses the rising and falling times of the pulses could affect the pulse shape. Even when an alternative is to use a *logic state analyser* (LSA), for our purposes we need to observe specific time stamps, where the rising and falling times of pulses are known, without the sampling time, more like in an analogue analysis. This enables to evaluate the coincidence time of pulses ($t_{0x}$). For these reasons, in this work we use the analogue analysis available in SPICE simulations.

For the measurement time observed in Fig. 2, in both cases the number of counted pulses and the number of obtained fractions are the same. But the measurand value is different; this characteristic changes the rate of occurrence of coincidences. In other words, even when the $P$ and $Q$ values increase at the same rate – because $f_{xs}$ and $f_{xp}$ have a difference of 4.6 ns, that is greater than the pulse width [16] – the corresponding fractions ($P_n/Q_n$) have different values. According to the classic theory of time-frequency metrology [11], a better approximation to the measurand will be obtained if the measurement time is increased.

The measurement of $f_{xs}$ and $f_{xp}$ is done using the same frequency standard ($f_0$). In consequence, the corresponding measurement time is given by:

$$M_{ts} = Q_{ns} T_0 = \frac{Q_{ns}}{f_0}, \tag{8}$$

$$M_{tp} = Q_{np} T_0 = \frac{Q_{np}}{f_0}, \tag{9}$$

respectively with sub-indexes *s* and *p* for the starting and posterior frequency values.

The fundamental assumption of the principle of rational approximations is the existence of coincidences. If two regular pulse trains are continuously compared, a coincidence pulse train is generated. In the literature there is reported that such a comparison requires the use multiplication of functions modelling each signal (in Fig. 1, the functions modelling $S_x$, $S_0$) [16]. The main novelty of this work is focused on simultaneous comparison of three signals: the original signal without a frequency shift ($S_{xs}$), the reference signal ($S_0$), and the signal after an unknown frequency shift ($S_{xp}$), where $f_s$, $f_0$, $f_{xp}$ are their respective frequency values.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 45–56.

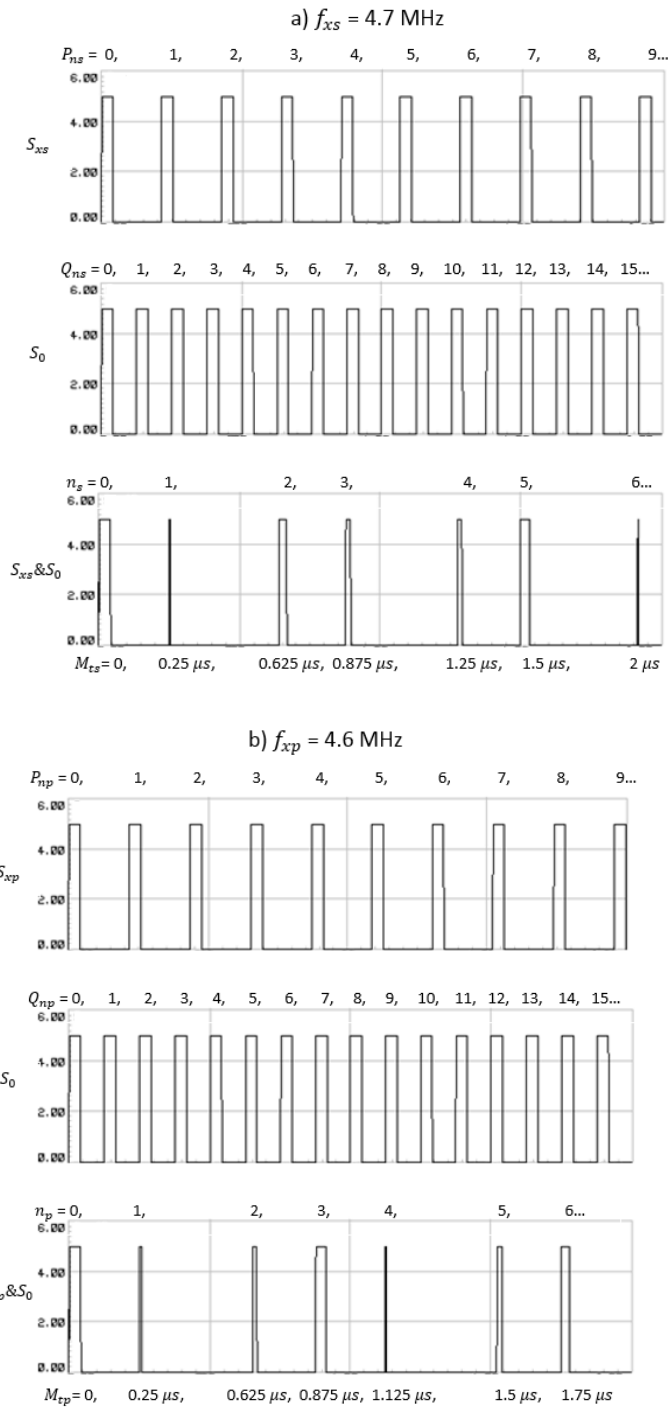



Fig. 2. Comparison of signals during the frequency measurement process, where $f_{xs}$ = 4.7 MHz, $f_{xp}$ = 4.6 MHz, $f_0$ = 8 MHz, and $\tau \approx 40$ ns.

Any signal comparison process has its measurement time associated with specific values of ($f_x$, $f_0$, $\tau$). From (8), (9):

$$M_{ts} = Q_{ns} \frac{M_{tp}}{Q_{np}}, \tag{10}$$

if $M_{ts} = M_{tp}$, then $Q_{ns} = Q_{np}$. The last relationships have sense, if we understand that the signal coincidence process $T_0$ is constant in $S_{xs}$ & $S_0$ and $S_{xp}$ & $S_0$. In consequence, during measurement in simultaneous comparison of three signals, the $P$-values ($P_{ns}$, $P_{np}$) are the only "truly" independents in continuous pulse counting (Fig. 3).

In the principle of rational approximations, after the first coincidence (when $M_t = 0$) of pulses from input signals, all the future coincidences are approximations to the measurand. When there is a comparison of three input signals, the measurement process starts when there is a coincidence of pulses from three input signals ($S_{xs}$ & $S_0$ & $S_{xp}$). It is worth remembering that for proper functioning of the principle of rational approximations, in order to have the same pulse width, all input signals must be conditioned. Following functioning of the principle of the rational approximations, after the $n$ – coincidence, the next, $(n+1)$ – coincidence is an approximation to the measurand. But in this case, rather than measuring the frequency value of a signal, the difference between frequencies of two signals ($S_{xs}$, $S_{xp}$) is measured. Basing on the previous analysis, the measured frequency difference between two signals ($S_{xs}$, $S_{xp}$) is defined as the frequency shift. (5) becomes:

$$\Delta f = f_{xs} - f_{xp}. \tag{11}$$

Since this proposal aims to measure the frequency shift occurring in a sensor's output, two sensors are required. Ideally, if two sensors of the same kind and operation range were identically fabricated – an example is measurement of the frequency shift from a QCM while it is loaded [18, 19] – they would have the same frequency value before a stimulus. The frequency of one sensor without a stimulus is the starting frequency ($f_{xs}$). On the other hand, the sensor frequency after stimulation is its posterior frequency ($f_{xp}$). Based on the last statements, an approximation to measurement of the frequency shift can be performed by using two sensors. One of them will be not stimulated, whereas the other one will be under a stimulus, which leads to a meaningful frequency shift. This approach is similar to the idea of differential measurement [21].

If the frequency values of starting and posterior frequencies are measured, after the first coincidence of pulses from the three input signals, they can be used for calculating the frequency shift:

$$\Delta f = \frac{P_{ns}}{Q_{ns}} f_0 - \frac{P_{np}}{Q_{np}} f_0, \tag{12}$$

$$\Delta f = \left( \frac{P_{ns}}{Q_{ns}} - \frac{P_{np}}{Q_{np}} \right) f_0 \tag{13}$$

for any $n > 0$. According to the principle of rational approximations, after the first coincidence the pulses of input signals are continuously counted. Any coincidence of the pulses of three input signals ($S_{xs}$, $S_{xp}$, $S_0$), is an approximation to $\Delta f$.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 45–56.

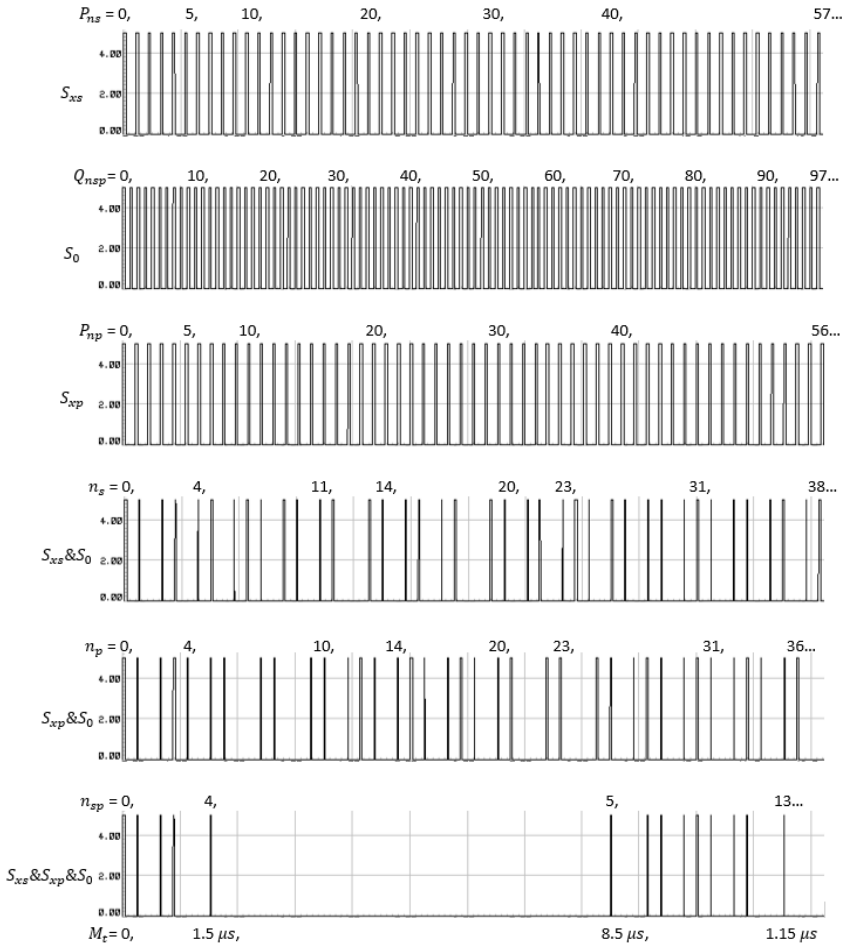



Fig. 3. Simultaneous comparison of three signals.

Since the reference frequency is the same when measuring both $f_{xs}$, $f_{xp}$:

$$\frac{f_{xs}}{f_0} - \frac{f_{xp}}{f_0} = \frac{P_{ns}}{Q_{ns}} - \frac{P_{np}}{Q_{np}}, \tag{14}$$

there is a value in the measurement time, where $Q_{ns} = Q_{np}$. At this moment, measurement of $f_{xs}$ and $f_{xp}$ can be done for calculating the corresponding frequency shift ($\Delta f$). (10) and (14) show a property of the physical phenomenon of signal comparison. For the ratios $f_{xs}/f_0$, $f_{xp}/f_0$ there is a measurement time where $M_{tp} = M_{ts}$. Knowing $f_0$, when $M_{tp} = M_{ts}$ counts of the reference pulses in both comparison processes give $Q_{ns} = Q_{np}$. This property explains why simultaneous comparison of three input signals enables to measure the signal frequency immediately after the shift. Such a behaviour is illustrated in Fig. 3. The generality of this formalism is illustrated by (14). For any given value of $f_{xs}$, $f_{xp}$ there is a measurement time, during which the frequency shift is calculated.

The nature of signal comparison enables to understand the effect of comparing three signals. It is mainly reflected in the measurement time needed for achieving measurement results.

As stated before, measurement of $\Delta f$ requires measuring separately $f_{xs}$ and $f_{xp}$. These operations require at least a total measurement time of:

$$M_{tsp} = M_{tp} + M_{ts} + O_t, \tag{15}$$

where the operation time ($O_t$) defines the time required for switching from measurement of $f_s$ to $f_p$. $O_t$ includes the time used by the operator of the frequency measurement instrument, or – in the case of automatic switching – the time required by the microcontroller. With the herein proposed approach the time required for measuring a frequency shift is given by:

$$M_{tfs} = Q_{ns} Q_{np} T_0. \tag{16}$$

The last equation shows the principal strength of the presented formalism. For the case of $M_{tsp}$, both signals ($f_{xs}$, $f_{xp}$) need to have enough time for being approximated, which depends directly on $Q_{ns}$, $Q_{np}$. In the traditional approach to the principle of rational approximations, $Q_{ns} \neq Q_{np}$. On the other hand, since $M_{tfs}$ uses simultaneous comparison of three signals, the $Q-$values increment at the same rate; in other words: $Q_{ns} = Q_{np} = Q_{nsp}$, where $nsp$ is the number of coincidence of three pulses of $S_{xs}$, $S_{xp}$, $S_0$, and $Q_{nsp}$ is the count value of the pulses in $f_0$ for coincidences $S_{xs}$ & $S_0$ & $S_{xp}$. One implication of these statements is that $M_{tfs} < M_{tsp}$. The last property shows how simultaneous comparison of three signals enables to obtain a better approximation than that of two signals at a different time. As a result, the frequency shift is calculated from (13), for any coincidence $nsp > 0$, by:

$$\Delta f = \left( \frac{P_{ns} - P_{np}}{Q_{nsp}} \right) f_0 \ , \tag{17}$$

and the measurement time is defined as:

$$M_t = M_{tfs} = Q_{nsp} T_0. \tag{18}$$

As it has been shown in previous works [12–16], the time required for measuring depends on overlapping of existing pulses at the same time. The likelihood of two overlapped pulses (generated independently) is greater than overlapping of three pulses, which are also generated independently. This leads to a longer time required for measuring than that for three signals.

In this Section, it was proposed a modification of the principle of rational approximations for application to measurement of frequency shifts. The next section is devoted to evaluating this theory.

## 4. Evaluation of measurement of frequency shifts

Measurement of frequency shifts uses simultaneous measurement of three signals, and – when the pulses of three signals are coincident - an approximation to the measurand is obtained (in this case $\Delta f$). In this section, using Matlab, we evaluate the formalism proposed in Section 3.

According to (13), the measurement of $\Delta f$ requires application of the principle of rational approximations for approximating the values of $f_{xs}$ and $f_{xp}$. In Fig. 4, after using the (1) – (4), a relationship between the relative error and the measurement time is shown for $f_{xs} = 4.7$ MHz – the data of Fig. 1. The duration of coincidences ($t_{0x}$) – when $f_{xs} = 4.7$ MHz – is illustrated in Fig. 2a. This shows different $t_{0x}-$values, which affect changing $\beta$, and – since counting of the pulses in $S_{xs}$ is continuous – $\beta$ decreases. According to (1), the best approximation occurs for $P_{ns}/Q_{ns} = 47/80$ (Fig. 3) or $M_{ts} = 1 \times 10^{-5}$ s (Fig. 4).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 45–56.

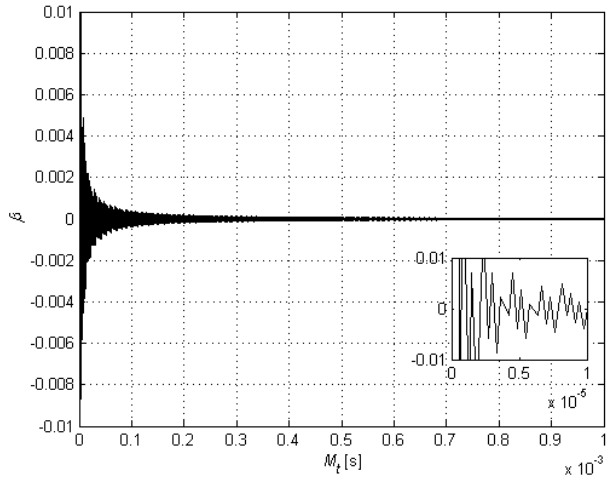



Fig. 4. The relative error in the frequency measurement process for $f_{xs} = 4.7$ MHz.

On the other hand, Figs. 5a and 5c show the relative error when measuring $f_{xp} = 4.6$ MHz and $f_{xp} = 4.699$ MHz. For the cases presented in Fig. 2 and Fig. 3, after $\Delta f = 100$ kHz, there are less pulses in $S_{xp}$ & $S_0$ than in $S_{xs}$ & $S_{xp}$. This effect can be observed only after a long measurement time; this fact is illustrated by comparing Fig. 2 and Fig. 3. The reason of this decrement – in the number of coincidences – is that the time difference in periods of signals ($T_{xs} - T_0$ or $T_{xp} - T_0$) becomes greater. The relative error in the measurement process is related to the duration of coincidence; if $t_{0x} = \tau$, there is the perfect coincidence and the best approximation to the measurand [16]. In a continuous signal comparison process, there are both partial ($t_{0x} < \tau$) and perfect coincidences, the difference is in the way these coincidences appear (Figs. 1–3). This phenomenon affects the way of decreasing $\beta$, and it is illustrated in the zoomed boxes in Fig. 4 and Fig. 5a. If the frequency shift is smaller, there are less coincidences. An example is when $\Delta f$ changes from 100 kHz to 10 kHz or $f_{xp}$ changes from 4.6 MHz (Fig. 5a) to 4.699 MHz (Fig. 5c). After evaluating both $\Delta f -$ values, the perfect coincidences exist in $P_{np}/Q_{np} = 23/40$ for $f_{xp} = 4.6$ MHz, and in $P_{np}/Q_{np} = 4699/8000$ for $f_{xp} = 4.699$ MHz. As we know, for obtaining the accurate approximation to the measurand during measurement, $f_x$ and $f_0$ must be stable within the measurement period – for the principle of rational approximations, this enables to obtain results comparable to those for the Allan variance [22]. The two fractions corresponding to the best approximations after the frequency shift require a measurement time of $5 \times 10^{-6}$ s for $P_{np}/Q_{np} = 23/40$ and $1 \times 10^{-3}$ s for $P_{np}/Q_{np} = 4699/8000$, and such calculations are true only for the stationary values of $f_{xp}$.

The approach proposed in this work enables to measure the frequency shift "when it occurs"; the last was analysed in the previous section, and it is obtained only by simultaneous comparison of three signals $S_{xs}$ & $S_0$ & $S_{xp}$. In this comparison, differences in periods of input signals ($S_{xs}$, $S_{xp}$ and $S_0$) – after their pulses coincide in time – generate different $P$, $Q -$ values, which can be used for calculating $\Delta f$ using (17). This is the main reason we could call this technique the principle of rational approximations for differential frequency measurements.

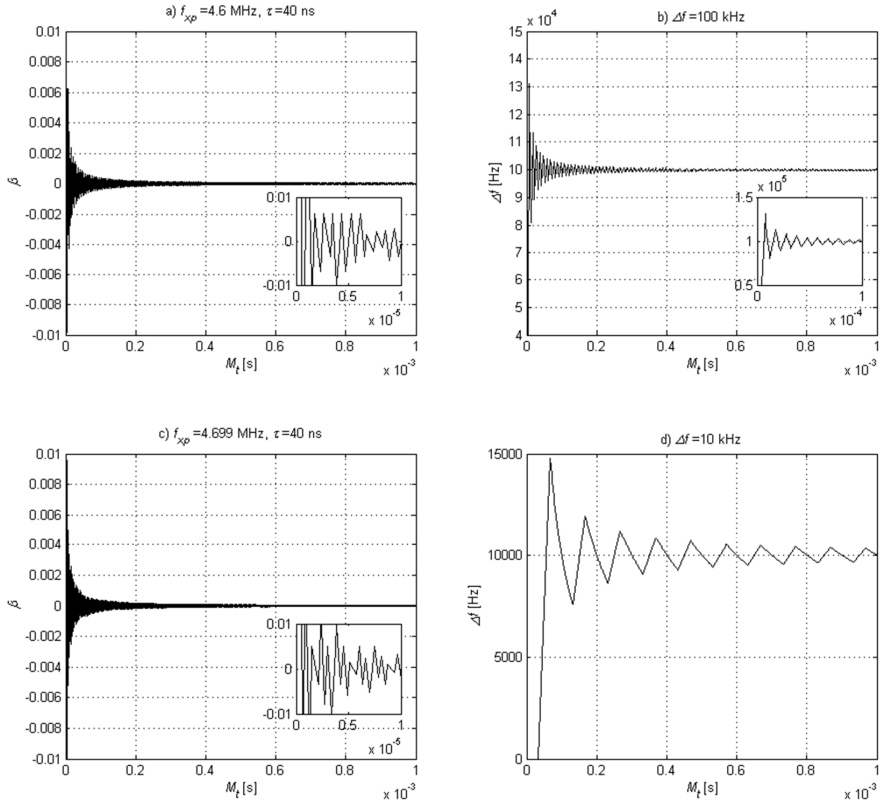



Fig. 5. The relative error in the frequency measurement process for $f_{xp}$ = 4.6 MHz (a); $f_{xp}$ = 4.699 MHz (c). Measurement of the frequency shift $\Delta f$=100 kHz (b); $\Delta f$=10 kHz (d).

In each coincidence ($S_{xs}$ & $S_0$ & $S_{xp}$), the frequency shift value is calculated using (13). For the examined cases, the value of $\Delta f$ is obtained in less than 1 ms. But another interesting property of our formalism is its clarity. The density of approximations is quite different for measuring $\Delta f$ and for any "single" measurement of $f_{xs}$ and $f_{xp}$. This phenomenon can be explained using Fig. 3. By comparing the number of pulses in $S_{xs}$ & $S_0$, $S_{xp}$ & $S_0$ with that in $S_{xs}$&$S_0$&$S_{xp}$, the variations in the number of coincident pulses is observed. Particularly, the lowest number of coincident pulses is observed in $S_{xs}$&$S_0$&$S_{xp}$. This can be explained by the fact that comparison of three signals leads to a lower number of coincidences. In consequence, less approximations are observed when measuring $\Delta f$.

After the analysis shown in this section, it is clear that the principle of rational approximations for differential frequency measurements is a novel technique for application to frequency sources with dynamic values (in particular FDS), where the measurement time required for obtaining an approximation to the measurand is shorter than that in the traditional principle of rational approximations.

## 5. Conclusions

Measurement of frequency shifts is required for sensors with high sensitivity. Well known frequency measurement techniques require more time for measuring if a greater accuracy is needed.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 45–56.

In this work, the theory of rational approximations is expanded by introduction of a new formalism for measuring frequency shifts. As a result, the principle of rational approximations is applied to differential frequency measurements, where $\Delta f$ can be measured in a very short time, without diminishing its accuracy. The proposal of this work is to measure frequency shifts by comparing three signals simultaneously. (10) – (12) showed that such a comparison is possible. Also, a further analysis – presented in Section 4 – illustrated how the coincidence of pulses occurs when comparing three signals, in a similar way like in the "traditional" principle of rational approximations, where only two signals are compared.

Our analysis shows that – due to the phenomenon of coincidence of signals – the measurements could be improved by choosing an optimal frequency standard, as well as a pulse width value. This analysis can be easily implemented using the algorithms of [16] and (10) – (12) and (17), (18).

## References


[1] Filippov, P., Strizhak, P.E., Vlasenko, N.V., Kochkin, Y.N., Serebrii, T.G. (2014). Adsorption-Desorption Dynamics of Alcohols on H-Beta and H-CMK Zeolites Nanocrystallites Studied by Quartz Crystal Microbalance Method. *Adsorption Science & Technology*, 32(10), 807–820.

[2] Afzal, N., Iqbal, A., Mujahid, Schirhagl. (2013). Advanced vapor recognition materials for selective and fast responsive surface acoustic wave sensors: A review. *Analytica Chimica Acta*, 787, 36 – 49.

[3] Arnau, A. (2008). Review of Interface Electronic Systems for AT-cut Quartz Crystal Microbalance Applications in Liquids. *Sensors*, 8(1), 370–411.

[4] Casteleiro-Roca, J.L., Calvo-Rolle, J.L., Meizoso-Lopez, M.C., Piñón-Pazos, A., Rodríguez-Gómez, B.A. (2014). New approach for the QCM sensors characterization. *Sensors and Actuators A: Physical*, 207, 1–9.

[5] Ishii, R., Naganawa, R., Nishioka, M., Hanaoka, T. (2013). Microporous organic-inorganic nanocomposites as the receptor in the QCM sensing of toluene vapors. *Analytical sciences: the international journal of the Japan Society for Analytical Chemistry*, 29, 283–289.

[6] Bhasker Raj, V., Singh, H., Nimal, A.T., Tomar, M., Sharma, M.U., Gupta, V. (2013). Effect of metal oxide sensing layers on the distinct detection of ammonia using surface acoustic wave (SAW) sensors. *Sensors and Actuators B: Chemical*, 187, 563–573.

[7] Kikuchi, M., Shiratori, S. (2005). Quartz crystal microbalance (QCM) sensor for $CH_3SH$ gas by using polyelectrolyte-coated sol-gel film. *Sensors and Actuators B: Chemical*, 108(1), 564–571.

[8] Bein, T., Mo, S., Mintova, S., Valtchev, V., Schoeman, B., Sterte, J. (1997). Growth of silicalite films on pre-assembled layers of nanoscale seed crystals on piezoelectric chemical sensors. *Advanced Materials*, 9(7), 585–589.

[9] Kirianaki, N.V., Yurish, S.Y., Shpak, N.O. (2001). Methods of dependent count for frequency measurements. *Measurement*, 29(1), 31–50.

[10] Kalisz, J. (2003). Review of methods for time interval measurements with picosecond resolution. *Metrologia*, 41(1), 17.

[11] Johansson, S. (2005). New frequency counting principle improves resolution. *Frequency Control Symposium and Exposition. Proc. of the 2005 IEEE International*, 628–635.

[12] Sergiyenko, O., Hernandez Balbuena, D., Tyrsa, V., Rosas Mendez, P.L.A., Rivas Lopez, M., Hernandez, W., Podrygalo, M., Gurko, A. (2011). Analysis of jitter influence in fast frequency measurements. *Measurement*, 44(7), 1229–1242.

[13] Murrieta-Rico, F.N., Mercorelli, P., Sergiyenko, O.Y., Petranovskii, V., Hernández-Balbuena, D., Tyrsa, V. (2015). Mathematical modelling of molecular adsorption in zeolite coated frequency domain sensors. *IFAC PapersOnLine*, 48(1), 41–46.

[14] Sergiyenko, O.Y. (2016). The mediant method for fast mass/concentration detection in nanotechnologies. *International Journal of Nanotechnology*, 13(1–3), 238–249.

[15] Hernandez Balbuena, D., Sergiyenko, O., Tyrsa, V., Burtseva, L., Rivas Lopez, M. (2009). Signal frequency measurement by rational approximations. *Measurement*, 42(1), 136–144.

[16] Murrieta-Rico, F.N., Yu, O., Sergiyenko, Petranovskii, V., Hernandez Balbuena, D., Lindner, L., Tyrsa, V., Rivas-Lopez, M., Nieto-Hipolito, J.I., Karthashov, V.M. (2016). Pulse width influence in fast frequency measurements using rational approximations. *Measurement*, 86, 67–78.

[17] Jansson, P.A. (1998). *Deconvolution of Images and Spectra*. New York: J. Wiley & Sons.

[18] Yu, L., Ma, X., Wu, T., Ma, Y., Shen, D., Kang, Q. (2016). Monitor the Processes of Ice Film Disappearance under a Stimulant Convection Condition and Absorption Ethanol Vapor to Ice by a Quartz Crystal Microbalance. *International Journal of Electrochemical Science*, 11(4), 2595–2611.

[19] Sasaki, I., Tsuchiya, H., Nishioka, M., Sadakata, M., Okubo, T. (2002). Gas sensing with zeolite-coated quartz crystal microbalances-principal component analysis approach. *Sensors and Actuators B: Chemical*, 86(1), 26–33.

[20] Sauerbrey, G. (1959). Verwendung von schwingquarzen zur wägung dünner schichten und zur mikrowägung. *Zeitschrift für physik*, 155(2), 206–222.

[21] Boyes, W. (2009). *Instrumentation Reference Book*. Butterworth-Heinemann.

[22] Allan, D.W., Barnes, J.A. (1981). A modified "Allan variance" with increased oscillator characterization ability. *Thirty Fifth Annual Frequency Control Symposium IEEE*.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 57–68.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# 5 PS JITTER PROGRAMMABLE TIME INTERVAL/FREQUENCY GENERATOR

**Paweł Kwiatkowski, Krzysztof Różyc, Marek Sawicki, Zbigniew Jachna, Ryszard Szplet**

*Military University of Technology, Faculty of Electronic, Gen. Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland*
*(✉ pawel.kwiatkowski@wat.edu.pl, +48 261 839 602, krzysztof.rozyc@wat.edu.pl, marek.sawicki@wat.edu.pl, zbigniew.jachna@wat.edu.pl, ryszard.szplet@wat.edu.pl)*

**Abstract**

A new time interval/frequency generator with a jitter below 5 ps is described. The time interval generation mechanism is based on a phase shifting method with the use of a precise DDS synthesizer. The output pulses are produced in a Spartan-6 FPGA device, manufactured by *Xilinx* in 45 nm CMOS technology. Thorough tests of the phase shifting in a selected synthesizer are performed. The time interval resolution as low as 0.3 ps is achieved. However, the final resolution is limited to 500 ps to maximize precision. The designed device can be used as a source of high precision reference time intervals or a highly stable square wave signal of frequency up to 50 MHz.

Keywords: time interval generator, digital-to-time converter, DDS synthesizer, phase shifting, FPGA.



## 1. Introduction

Generation of precise *time intervals* (TI) is a complementary process to their measurement. Both time interval counters and *time interval generators* (TIG) are widely used in many industrial and scientific areas. In addition, the identification of parameters of time interval counters depends on the quality of calibration and measurement procedures, that are performed using reference TIs with the aid of TIGs [1–3].

Short TIs can be generated using *e.g.* cables of different lengths, a digital-to-analogue converter and a ramp generator [4–6], delay lines [7–9] or DLLs [10, 11]. To obtain a wider generation range an interpolation method, taken from time interval counters, may be applied [12]. Using this approach the TI is divided into coarse and fine parts. The first one is obtained by counting periods of a reference clock while the second one is based on one of previously mentioned generation methods. A commonly used interpolation method in TIGs involves phase shifting of a reference clock [3, 13–16]. Other ways of getting both wide generation range and high TI resolution are based on counting a certain number of periods of selected values [17] or finding coincidence between two clocks that operate on slightly different frequencies [18].

Along with the development of digital electronics TIGs are willingly implemented using FPGA chips. Programmable arrays can be used either as a complete platform that performs timing generation [3, 15, 16, 18] or just as a part of generator [13, 14, 17, 19, 20]. FPGA build-in DLLs and PLLs greatly facilitate TIG integration in a single chip but, so far, cannot generate as low-jitter TIs as those using external modules, *e.g.* a precise DDS synthesizer [13, 21]. Furthermore, modern DDS chips achieve phase shifting of the output clock with a sub-picosecond step that is still difficult to obtain in FPGA.

Based on promising results obtained in [13] we designed a new TIG that employs a DDS synthesizer for fine phase shifting and an FPGA chip for coarse period counting and pulse generation. Special attention was given to characterization of DDS synthesizer features regarding phase shifting.

Article history: received on Jul. 20, 2016; accepted on Sep. 29, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0009.

## 2. Generator design

A block diagram of the designed TIG is shown in Fig. 1. The reference clock signal of 10 MHz frequency provided by either the internal *oven-controlled crystal oscillator* (OCXO) or by any external reference source (*e.g.* rubidium standard), is used to create a high frequency clock signal of 1 GHz in a DSPLL synthesizer. Due to the advantage of DSPLL technology the generated signal has an ultra-low jitter and a high resistance to *process, voltage and temperature* (PVT) variations [22]. Based on this signal the DDS synthesizer produces the main clock signal that can be precisely tuneable regarding the phase shift and frequency. The discrete sample values generated at the synthesizer output are filtered in the *low-pass filter* (LPF) to get a sinusoidal signal. Then the fast discriminator is used to obtain a square wave clock. The clock signal is fed to the FPGA device where output pulses of the TIG are generated by the pulse selector and frequency divider. The pulse generation is initialized with regard to the trigger signal that can be produced internally by the trigger generator or with the use of an external signal. The high quality of clock signals is guaranteed by the use of ultra-low jitter *low-voltage positive-referenced emitter-coupled logic* (LVPECL) standard devices applied in the clock signal path. It applies particularly to the DDS synthesizer, which is characterized by a very low value of residual phase noise (up to −152 dBc/Hz [23]).

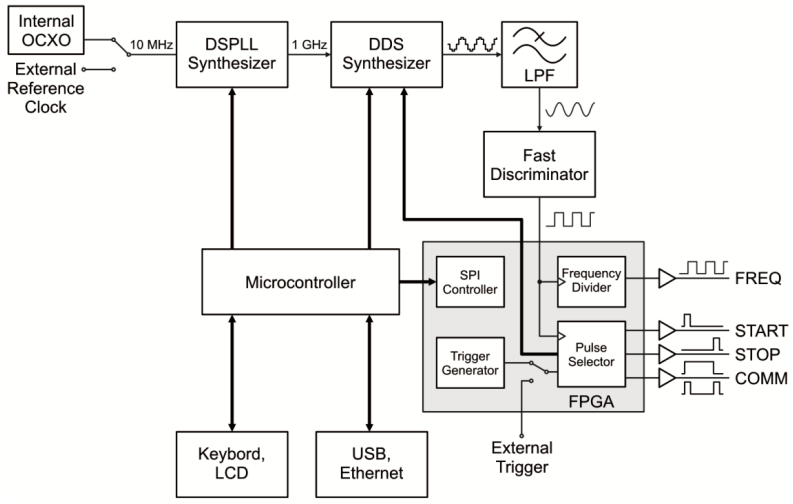



Fig. 1. A block diagram of the TIG.

The TIG can operate either in the time interval mode (START, STOP or COMM outputs) or in the frequency mode (FREQ output). In the first one the TI can be represented as (1) a pair of pulses at two separate outputs (START and STOP), (2) a pair of pulses at a single output (COMM, the time interval common mode), or (3) a single pulse with a declared width (COMM, the time interval width mode). The frequency of the main clock signal in the time interval mode is 50 MHz. The phase shifting of a user-selected value $\Delta\phi$ is performed by the DDS synthesizer synchronously with the START pulse generation in the pulse selector. The STOP pulse is generated after elapsing $N$ periods of the main clock signal and the time interval corresponding to $\Delta\phi$. In the frequency mode the main clock signal for the frequency divider is precisely selected in the DDS within a range from 50 MHz to 100 MHz. The divider enables to further expand the frequency generation range. The pulses generated pulses from the FPGA are distributed (in the low-voltage differential signaling standard LVDS) to the output buffers

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 57–68.

to obtain possibly steep slopes (rise and fall times < 250 ps) at 50 Ω load. Control of the device is provided either remotely via USB or Ethernet interfaces, or locally with the use of the LCD panel and keyboard integrated with the TIG. All user settings are transformed by the microcontroller into adequate executing commands for the DDS synthesizer and logic implemented in the FPGA chip, *i.e.* the pulse selector, frequency divider and trigger generator.

The designed TIG is packed into a Rack 2U case (Fig. 2) that facilitates its use in laboratory conditions. The front panel provides output signals and a local interface. Inputs of an external trigger and a reference clock are placed on the rear panel together with remote interfaces.


Fig. 2. An external view of the TIG.

### *2.1. Functional blocks implemented in FPGA chip*

FPGA are high performance devices; they operate on relatively high frequency clock signals and can include many user-programmable logic circuits that work autonomously. Hence, to generate the output pulses we selected a Spartan-6 FPGA device manufactured by *Xilinx*. Four modules, *i.e.*: the pulse selector, frequency divider, trigger generator and SPI controller (Fig. 1) were designed using the hardware description language VHDL and were implemented in the FPGA device with the aid of the ISE Design Suite firmware environment [24]. The main task of the FPGA is to generate two output pulses that are mutually shifted in time. This task is executed in the pulse selector module presented in Fig. 3.

The trigger signal, after synchronization to the main clock ($CLK_{\mathrm{DDS}}$) in the double synchronizer, initializes phase shifting in the DDS synthesizer. This operation has to be performed with regard to the DDS synchronization clock ($CLK_{\mathrm{SYNC}}$) of 250 MHz frequency. Thus, another double synchronizer is used for this clock. The DDS profile selector switches between two preselected DDS synthesizer configurations: one with and one without phase shifting. The switching process takes slightly above 100 ns, after execution of the phase shift. To ensure generation of a time interval shorter than 100 ns an additional phase shift delayer is used. It contains a set of serially connected flip-flops. Finally, the delayer initializes generation of START and WIDTH pulses and counting periods in the modulo $N$ counter on the rising edge of the last period of $CLK_{\mathrm{DDS}}$ signal that does not contain phase shifting. Then configuration switching is executed in the DDS synthesizer and the main clock signal is shifted in time by a preselected value $\Delta\phi$. The counter operates until it reaches a declared value of $N$. That causes resetting the double synchronizer and phase shift delayer. Finally, the STOP pulse and falling edge of the WIDTH pulse are generated. The maximum value of $N$ can be $2^{29}$. Taking into account the period of $CLK_{\mathrm{DDS}}$ clock signal ($T_0 = 20$ ns) the maximum TIG range is slightly above 10 s. The pulse shapers are responsible for generation of pulses with a constant width (START, STOP) and for selection of the operation mode of COMM output (width or common). Also, polarization of pulses can be set.

The square wave signal of a desired frequency is generated in two steps. Firstly, the output frequency of the DDS synthesizer is precisely tuned in a limited range (50 MHz – 100 MHz). Then, the obtained signal is divided by the selectable frequency divider implemented in FPGA. In this way frequencies within an operation range from 0.1 Hz to 1 MHz can be chosen with 1 mHz step, while for the range from 1 MHz to 50 MHz – with a step of 1 Hz. Due to the fact that the pulse selector and the frequency divider operate on the same clock signal, both modules

cannot run in the same time. Therefore, either the time interval mode (active START, STOP and COMM outputs) or the frequency mode (active FREQ output) can be selected in a given time.

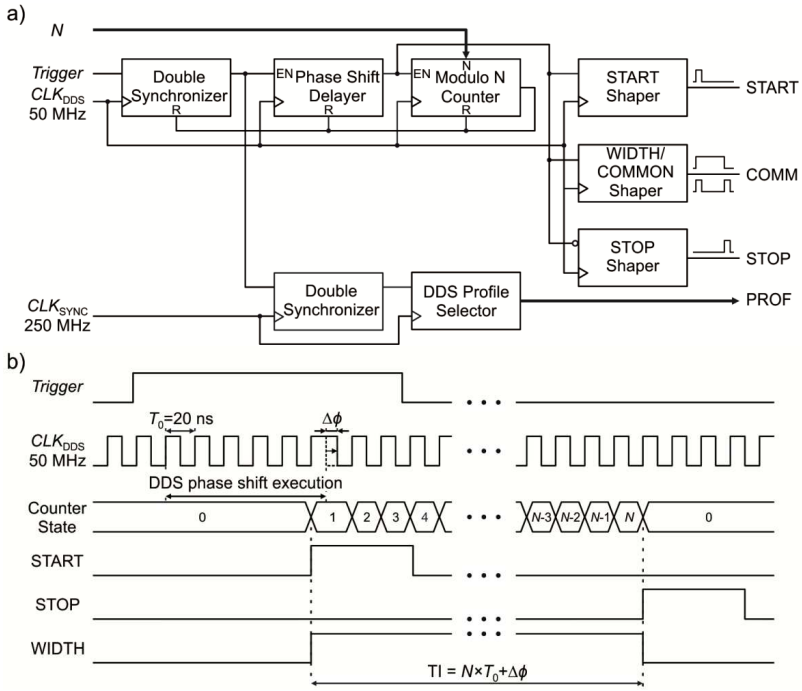



Fig. 3. A block diagram (a) and the operation principle (b) of the pulse selector.

The trigger generator is used to initialize the generation process in the pulse selector. It consists of several clock dividers that are fed by the asynchronous signal with regard to the $CLK_{DDS}$ clock. The SPI controller (Fig. 1) receives commands from the microcontroller, such as: "set *N*" for the modulo counter or "divide value" for the frequency divider, "choose an operation mode (time interval/frequency, width/common mode)", "select a trigger (internal/ external, internal trigger frequency)".

### *2.2. DDS Synthesizer features*

The TIG parameters depend mainly on the main clock quality and possibilities of its phase shifting. Therefore, the choice of synthesizers is crucial. The selected AD9910 (*Analog Devices*) DDS synthesizer is characterised by 0.23 Hz frequency resolution and $1/2^{16}$ clock period phase shift step [23]. If 50 MHz clock is applied, the minimum $\Delta\phi$ reaches a sub-picosecond value (20 ns / $2^{16} \approx 305$ fs).

We considered three possibilities of using DDS synthesizers. In the first approach we used two of them synchronized to each other. The first synthesizer generated a clock signal for the START shaper, whereas the second one – a signal of the same frequency but shifted in time for the STOP shaper. Both shapers were driven by separate signals so it was possible to generate time intervals from 0 duration. However, the synchronization of both synthesizers caused extension of the generated time interval jitter value above 20 ps. It is a reasonably good result in the case of synchronization, but it is insufficient for the expected TIG precision. For this reason further work with that design was discontinued.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 57–68.

The second and third approaches are based on a single DDS synthesizer but differ in the way of phase shift control. According to [23], the value of phase shift is set by the *Phase Offset Word* (POW) register. The value can be changed either by using a dedicated 16-bit parallel port or by choosing preselected profiles. Each profile consists of a group of 8 registers that contain operating parameters for the output signal. A particular profile is activated using a 3-bit PROF port. In the case of TIG two profiles can be configured as signals of the same frequency and amplitude but of different phase shift values. When profiles are changed then only the content of POW register is modified. The parallel port provides direct access to the POW register. In both cases – the parallel port and profile switching – a similar phase shift execution time is needed. Thus, we have chosen the profile selection that minimizes the number of connections on a PCB. Because switching is performed only between two profiles then only one bit of PROF port is needed.

## 3. Experimental tests

### *3.1. Test setup*

The TIG was examined in a test setup shown in Fig. 4. The generator was placed in a climatic chamber PL-2J (*ESPEC*). The short TIs between START and STOP pulses ($\leq$ 10 μs) were measured using a DSA90804A oscilloscope (*Keysight*) [25] and the longer ones with the use of a newly developed precise time interval counter with a precision even as low as 3 ps [26]. Both the TIG and the counter were driven by two separate rubidium generators FS725 (*Stanford Research Systems*).

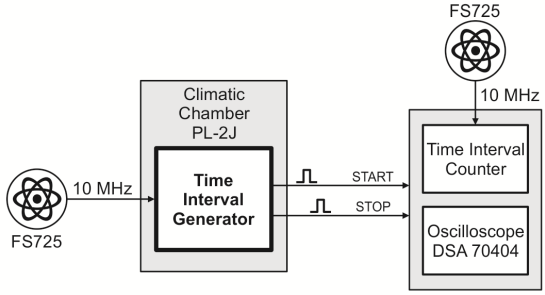



Fig. 4. A test setup for TIG.

### *3.2. In-period TIG evaluation*

Since the TIG performance is largely dependent on the DDS synthesizer, much work was devoted to precisely examine its features regarding phase shifting. We tested the transfer function and time jitter of TIs generated for all POW values. The research was done using a high performance oscilloscope that calculated mean value and standard deviation of TIs based on 1000 samples. Each measurement of the TI length and jitter lasted about 6 seconds. The obtained results for sine and cosine signals are presented in Fig. 5.

In both cases, a cyclic increase of the jitter value is observed. Looking carefully into the conversion characteristic it can be seen that the obtained phase shift step is smaller (the slope of blue line) in the same POW ranges where lower values of timing jitter occur. This feature may result from the process of counting samples in the DDS synthesizer.

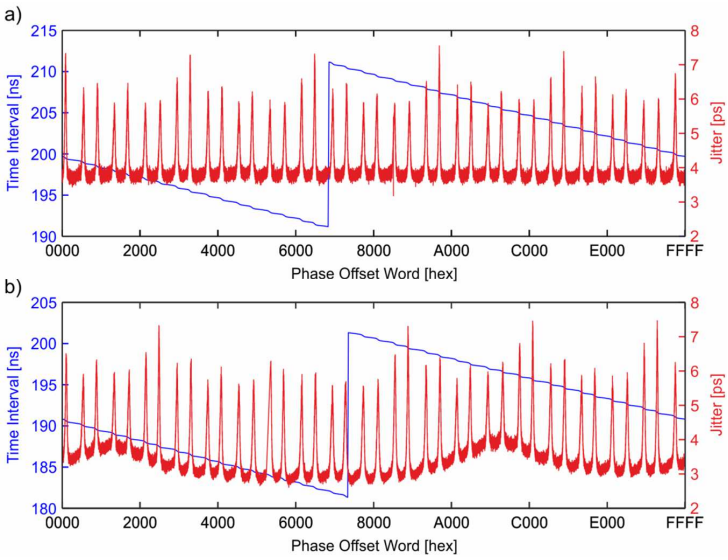



Fig. 5. The length of TI and its jitter within a single period of the main clock for a sine (a) and cosine signal (b).

Unequal length changes of the generated TIs for successive POW values (blue line, Fig. 5) lead to relatively high nonlinearities of conversion. It is common to define nonlinearities using two parameters, *i.e.* differential and integral nonlinearity (denoted by DNL and INL, respectively). The first one describes a difference between the actual size of the *i*-th phase shift step ($q_i$) and the mean step size ($q_m$, denoted also as the value of least significant bit LSB). It is expressed as:

$$\mathrm{DNL}_i[\mathrm{LSB}] = \frac{q_i}{q_m} - 1, \tag{1}$$

where the mean step size $q_m$ is calculated as a conversion characteristic range ($T_R$) divided by the number of steps ($M$). The second parameter (INL) defines how much the actual conversion characteristic differs from the ideal one in the *i*-th phase shift step. The corresponding equation is as follows:

$$\mathrm{INL}_j = \sum_{i=1}^{j} \mathrm{DNL}_i, \tag{2}$$

where $1 \le j \le M$.

Regarding the designed TIG $q_m$ (denoted as LSB) is equal to 305 fs ($T_R = T_0 = 20$ ns, $M = 2^{16} = 65536$). After reordering TI values, obtained for all POWs, in the increasing order and after calculating differences between the neighbouring values we obtain the actual step sizes $q_i$. The TIG nonlinearity, evaluated using (1) and (2), is presented in Fig. 6 as bar graphs of DNL and INL values. The maximum DNL is equal to about 31 LSB, which corresponds to 9.5 ps, while the extreme value of INL is 714 LSB, *i.e.* 218 ps.

For most POW values the time interval jitter is below 4 ps, while peaks can reach even 7.5 ps. The characteristics of sine and cosine signals are slightly different. The cyclic growths of time jitter for both signals partly overlap. Therefore, the jitter cannot by minimized by dynamic switching between the sine and cosine types of output signal.

Experimental tests to obtain all possible values of POW ($2^{16} = 65536$) for measurements of both sine and cosine signals lasted about 9 days. To check the reproducibility of measurement results the test was repeated. Fig. 7 shows the difference between the jitter values and the time

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 57–68.

interval lengths obtained in the first and the second tests for a sine signal. The time interval jitter in both series is within a range from –0.94 ps to 0.75 ps, while the length of generated TIs varies within a range from –9.59 ps to 21.14 ps. Assuming no temperature influence (the tested generator was placed in a climatic chamber that stabilized temperature at 21°C), the obtained slight differences are caused mainly by: (1) the stability of the reference 10 MHz clock signal, (2) the stability of the DSPLL synthesizer 1 GHz output signal, and (3) the quality of the main clock produced by the DDS synthesizer with dynamic phase shifting.

A common problem with integrated circuits is their vulnerability to PVT variation. The process variation is related to a method of CMOS chip fabrication [9]. Thus, some differences in performance of chips are acceptable. However, the use of digital techniques such as DSPLL and DDS makes the presented TIG resistant, to some extent, to the process variation. As a proof, we have tested the second TIG device and obtained very similar results as in Fig. 7. The environmental (power supply voltage and temperature) conditions can vary in time and space resulting in changes of the generated TI length and in the increased jitter [9]. The influence of voltage variation is limited in the designed TIG by applying multi-stage voltage stabilization using ultra low-noise low-dropout regulators, while the impact of temperature variation is discussed further in the paper (Subsection 3.3).

Due to inability to eliminate the cyclic growth of the time jitter of generated TIs we decided to reduce the TIG resolution to 0.5 ns. This enables to arbitrarily choose the phase shift values for which the time jitter of generated TIs does not exceed 4 ps (Fig. 8a). With such a limited resolution the INL error of the designed TIG reaches very low values with the extreme one equal to |–0.0013| LSB = 0.65 ps (Fig. 8b).

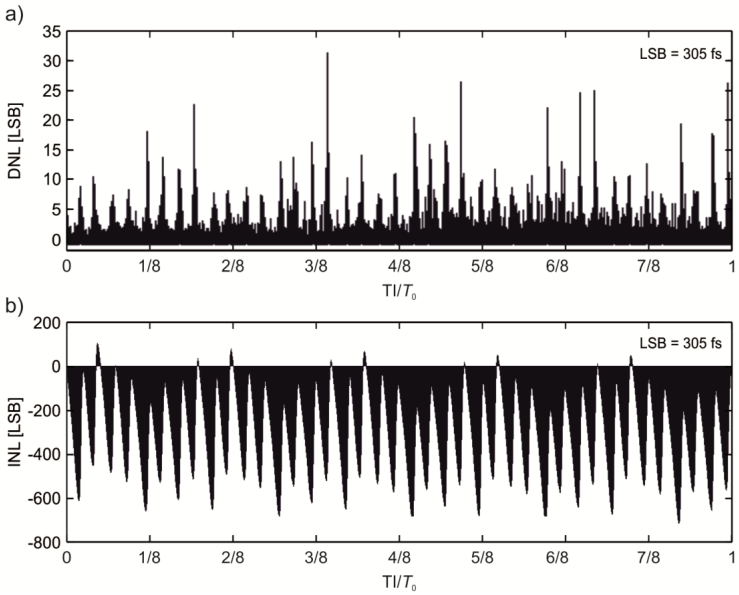



Fig. 6. The DNL (a) and INL (b) plots of the TIG with LSB = 305 fs.

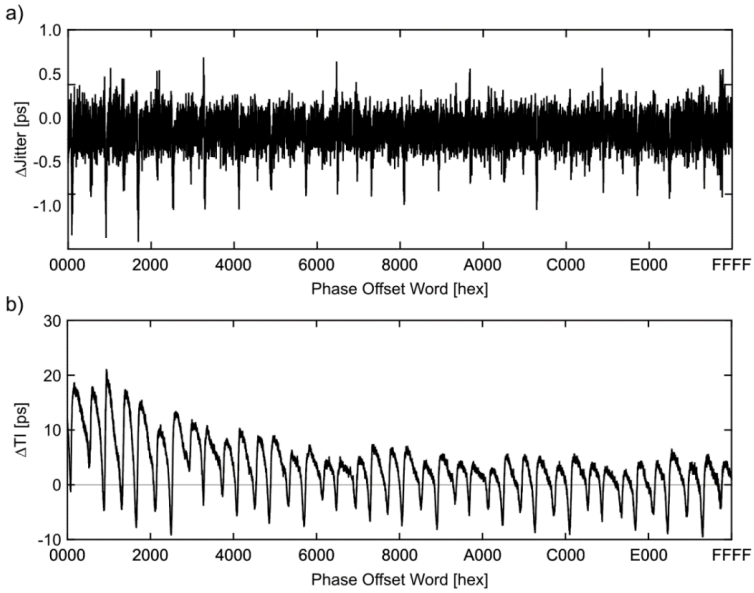



Fig. 7. Differences of jitter (a) and lengths of TI (b) for two independent series of measurements for a sine signal.

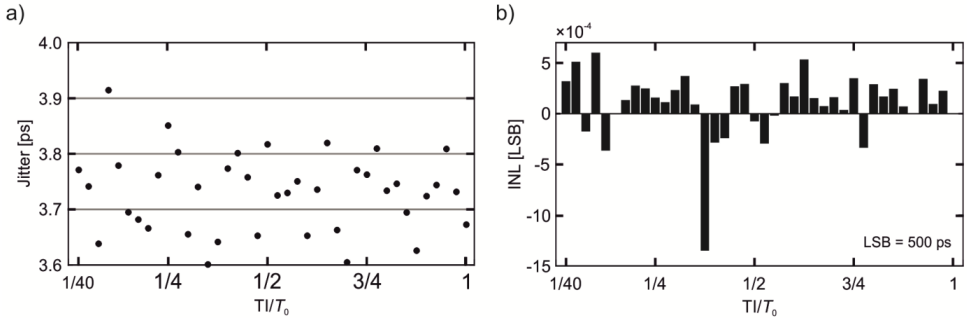



Fig. 8. The time interval jitter (a) and INL plot (b) of the TIG within the range of a single $T_0$ and resolution of 500 ps.

### *3.3. Wide range TIG evaluation*

In the next test the temperature inside the thermal chamber was stable (21°C) and measurements of the time interval jitter were performed for different TI lengths. As a reference clock we used either the internal built-in OCXO or an external rubidium generator. The results are shown in Fig. 9. The most precise measurements are done with the use of an oscilloscope. Thus, in a range of up to 10 μs the time jitter values are below 5 ps. The time interval counter has a greater intrinsic standard measurement uncertainty than the oscilloscope, so that the obtained jitter is a bit higher. The short-term stability of the reference clock becomes a critical parameter for generated TIs exceeding 20 ms. A more stable clock (*i.e.* a rubidium generator) enables to keep the jitter below 10 ps within a range of up to 50 ms.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 57–68.

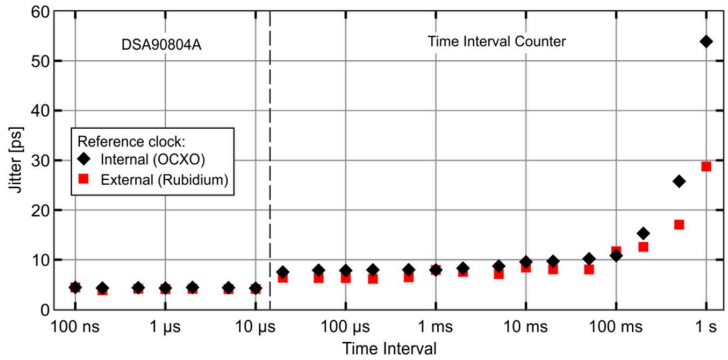



Fig. 9. The time interval jitter of generated TIs.

### *3.4. Temperature tests*

During the thermal tests the TI length was arbitrarily selected as 14,500 µs and the ambient temperature varied from −10°C up to +60°C. The measured value of TI length was changed less than 20 ps with regard to the value obtained at 20°C (Fig. 10). So the ambient temperature changes have a little influence on the TIG accuracy. Also the time interval jitter changes slightly, *i.e.* less than 0.8 ps. Contrary to Fig. 10a, there is no visible trend in Fig. 10b. It can be concluded that the obtained differences come from the noise floor of the oscilloscope input circuits [25]. Therefore, temperature changes do not affect the TIG precision.

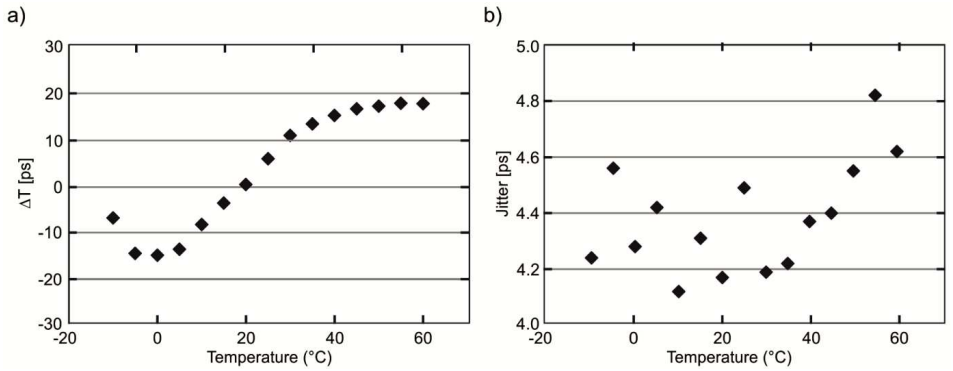



Fig. 10. The temperature dependence of the length of generated TIs (a) and the time interval jitter (b).

### *3.5. Frequency test*

In the last experiment we checked the frequency mode of the TIG. A square wave signal of a selected frequency was measured using an SR620 counter (*Stanford Research Systems*) driven by an external FS725 reference clock generator. Fig. 11 shows the example results of detuning for a frequency range from 10 Hz to 50 MHz. The detuning values may result from the volatility of the build-in OCXO reference signal source. Because the values of detuning are linearly dependent on the selected frequency, they can be easily compensated.

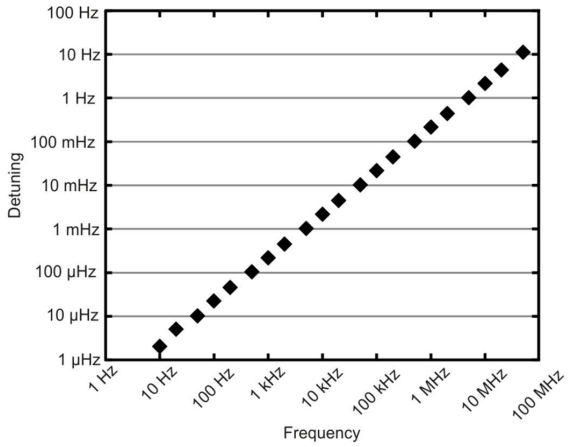



Fig. 11. The detuning for a frequency range from 10 Hz to 50 MHz.

Using a precise time interval counter the *Allan deviation* (ADEV) value of generated signal of 10 MHz frequency was evaluated. The obtained results are summarized below:
ADEV 0.1 s – $1.74\times10^{-9}$, 1 s – $2.42\times10^{-10}$, 10 s – $5.10\times10^{-11}$.

## 4. Conclusions

The designed generator produces time intervals with the time interval jitter less than 5 ps and the resolution of 500 ps. A low jitter is achieved thanks to the use of an ultra-low jitter synthesizer device. Selected DDS synthesizer features restrict the possibility of obtaining the time interval jitter below 5 ps together with a sub-picosecond resolution. The use of digital methods for generating time intervals makes the generator resistant to PVT variations. The generator can also produce a frequency signal of high stability (*e.g.* ADEV(1 s)= $2.42\times10^{-10}$).

## Acknowledgements

This work was supported by the Polish National Centre for Research and Development under contract no. PBS1/B3/3/2012

## References

[1] Rivoir, J., (2006). Full-digital time-to-digital converter for ATE with autonomous calibration. *Proc. of IEEE Int. Test. Conf. 2006*, Santa Clara, CA, USA, 1–10.

[2] Szplet, R., Jachna, Z., Kwiatkowski, P., Różyc. K. (2013). A 2.9 ps equivalent resolution interpolating time counter based on multiple coding lines. *Meas. Sci. Technol.*, 24(3), 035904/1–15.

[3] Vornicu, I., Carmona-Galan, R., Rodriguez-Vazquez, A. (2016). Time interval generator with 8 ps resolution and wide range for large TDC array characterization. *Analog. Integr. Cir. Sig. Process*, 87(2), 181–189.

[4] Using Digitally Programmable Delay Generators. AN-260 Application Note, Analog Devices. *http://www.analog.com/media/en/technical-documentation/application-notes/105895411AN-260.pdf.* (1998).

[5] Alhdab, S., Mantyniemi, A., Kostamovaara, J. (2012). A 12-bit Digital-to-Time Converter (DTC) with sub-ps-level resolution using current DAC and differential switch for Time-to-Digital Converter (TDC). *Proc. of IEEE I2MTC 2012*., Graz, Austria, 2668–2671.

[6] Klepacki, K., Pawłowski, M., Szplet, R. (2015). Low-jitter wide-range integrated time interval/delay generator based on combination of period counting and capacitor charging. *Rev. Sci. Instrum.*, 86(2), 025111/1–7.

[7] Rahkonen, T., Kostamovaara, J. (1993). The use of stabilized CMOS delay line for the digitization of short time intervals. *IEEE J. Solid-State Circuit*s, 28(8), 887–894.

[8] Suchenek, M. (2009). Picosecond resolution programmable delay line. *Meas. Sci. Technol.*, 20(11), 117005/1–5.

[9] Abdulrazzaq, B.I., Abdul Halin, I., Kawahito, S., Sidek, R.M., Shafie, S. Yunus, N.A.M. (2016). A review on high-resolution CMOS delay lines: towards sub-picosecond jitter performance. *SpringerPlus*, 5(1), 1–32.

[10] Huang, H.-Y., Shen, J.-H. (2004). A DLL-based programmable clock generator using threshold-trigger delay element and circular edge combiner. *Proc. of IEEE AP ASIC 2004*, Fukuoka, Japan, 76–79.

[11] Okayasu, T., Suda, M., Yamamoto, K., Kantake, S., Sudou, S., Watanabe, D. (2006). 1.83ps-resolution CMOS dynamic arbitrary timing generator for > 4 GHz ATE applications. *Proc. of IEEE ISSCC 2006*, San Francisco, CA, United States, 522–511.

[12] Carbone, P., Kiaei, S., Xu, F. (2014). *Design, modelling and testing of data converters*. Berlin, Germany: Springer-Verlag, ch. 7.

[13] Kwiatkowski, P., Jachna, Z., Różyc, K., Kalisz, J. (2012). Accurate and low jitter time-interval generators based on phase shifting method. *Rev. Sci. Instrum.*, 83(3), 034701/1–4.

[14] Suchenek, M., Starecki, T. (2012). Programmable pulse generator based on programmable logic and direct digital synthesis. *Rev. Sci. Instrum.*, 83(12), 124704/1–4.

[15] Chen, Y.-Y., Huang, J.-L., Kuo, T., Huang, X.-L. (2015). Design and implementation of an FPGA-based data/timing formatter. *J. Electron. Test.*, 31(5–6), 549–559.

[16] Yao, Y., Wang, Z., Lu, H., Chen, L., Jin, G. (2016). Design of time interval generator based on hybrid counting method. *Nucl. Instrum. Methods Phys. Res., Sect. A*, 832, 103–107.

[17] Kalisz, J., Poniecki, A., Różyc, K. (2003). A simple, precise, and low jitter delay/gate generator. *Rev. Sci. Instrum.*, 74(7), 3507–3509.

[18] Chen, P., Chen, P.-Y., Lai, J.-S., Chen, Y.-J. (2010). FPGA vernier digital-to-time converter with 1.58 ps resolution and 59.3 minutes operation range. *IEEE Trans. Circuits Syst. I, Reg. Papers*, 57(6), 1134–1142.

[19] Song, Y., Liang, H., Zhou, L., Du, J., Ma, J., Yue, Z. (2011). Large dynamic range high resolution digital delay generator based on FPGA. *Proc. of ICECC 2011*, Zhejiang, China, 2116–2118.

[20] Miari, L., Antonioli, S., Labanca, I., Crotti, M., Rech, I., Ghioni, M. (2015). Eight-channel fully adjustable pulse generator. *IEEE Trans. Instrum. Meas.*, 64(9), 2399–2408.

[21] Kwiatkowski, P., Szplet, R., Jachna, Z., Różyc, K. (2016). A time digitizer based on multiphase clock implemented in FPGA device. *Proc. of EBCCSP 2016*, Cracow, Poland.

[22] Optimizing clock synthesis in small cells and heterogeneous networks. White Paper, Silicon Laboratories. *https://www.silabs.com/Support%20Documents/TechnicalDocs/Silicon%20Labs%20Next-Generation%20DSPLL%20Technology%20White%20Paper%20-%20June%202015.pdf.* (Jun. 2015)

[23] 1 GSPS, 14-Bit, 3.3 V CMOS Direct Digital Synthesizer. Datasheet, Analog Devices. *http://www.analog.com/media/en/technical-documentation/data-sheets/AD9910.pdf.* (May 2012).

[24] ISE In-Depth Tutorial. User Guide UG695, v.14.1, Xilinx. *http://www.xilinx.com/support/documentation/sw_manuals/xilinx14_1/ise_tutorial_ug695.pdf.* (Apr. 2012).

[25] Keysight Technologies Infiniium 90000 Series Oscilloscopes. Datasheet, Keysight Technologies. *http://literature.cdn.keysight.com/litweb/pdf/5989-7819EN.pdf.* (2015).

[26] Szplet, R., Kwiatkowski, P., Jachna, Z., Różyc, K. (2016). An eight-channel 4.5-ps precision timestamps-based time interval counter in FPGA chip, *IEEE Trans. Instrum. Meas.*, 65(9), 2088–2100.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 69–78.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# ESTIMATION OF CONDITIONAL EXPECTED VALUE FOR EXPONENTIALLY AUTOCORRELATED DATA

**Adam Kowalczyk, Anna Szlachta, Robert Hanus, Rafał Chorzępa**

*Rzeszow University of Technology, Department of Metrology and Diagnostic Systems, Powstańców Warszawy 12, 35-959 Rzeszów, Poland (kowadam@prz.edu.pl, ✉ annasz@prz.edu.pl, +48 17 743 2462, rohan@prz.edu.pl, rchorz@prz.edu.pl)*

**Abstract**

Autocorrelation of signals and measurement data makes it difficult to estimate their statistical characteristics. However, the scope of usefulness of autocorrelation functions for statistical description of signal relation is narrowed down to linear processing models. The use of the conditional expected value opens new possibilities in the description of interdependence of stochastic signals for linear and non-linear models. It is described with relatively simple mathematical models with corresponding simple algorithms of their practical implementation. The paper presents a practical model of exponential autocorrelation of measurement data and a theoretical analysis of its impact on the process of conditional averaging of data. Optimization conditions of the process were determined to decrease the variance of a characteristic of the conditional expected value. The obtained theoretical relations were compared with some examples of the experimental results.

Keywords: conditional averaging, conditional expected value, auto-correlated data, random signals.



## 1. Introduction

Intense worldwide technological development poses new challenging tasks for metrology. In measurement models, relations between stochastic variables (signals) should be taken into account. Stochastic relations make it harder to estimate the statistical characteristics of signals and measurement data. The majority of existing documents and measurement recommendations intended for use do not take into account the impact of stochastic relations between data [1–4].

Familiarity with the characteristics which describe stochastic relations is the primary question in solving many issues in science and technology. The most commonly used characteristics of stochastic relations are correlation functions. The need to take data correlation into account in evaluation of the measurement result uncertainty is indicated by many authors in specialised publications. An analysis of time series in the form of auto-correlated numerical sequences is presented in the study [5]. Determination of the impact of autocorrelation through an alternative measure, the so-called effective number of independent observations, was undertaken in a number of publications [6–8]. The derivation, analysis and examples of the use of formulae on the unbiased single measurement variance estimators and the arithmetic mean for correlated data, and a discussion on metrological usability of the proposed characteristics were presented in the works [9–11]. Simulation research into the impact of the instability of estimates of a normalised autocorrelation function on the uncertainty of the arithmetic mean value was carried out using the Monte Carlo method in the paper [12]. The application of Allan's variance in the analysis of correlated data is shown in the paper [13].

The theoretical correlation characteristics, which give effective results with linear relations, lose their benefits in the analysis of signals and systems with nonlinear characteristics of relations. In practice, in measurement of data with stochastic relations, correlation characteristics tend to create computational difficulties. There is a need to determine

Article history: received on May 11, 2016; accepted on Sep. 19, 2016; available online on Mar. 15, 2017; DOI: 10.1515/mms-2017-0005.

the behaviour of the autocorrelation function and the sign of correlation. Moreover, the literature does not provide accurate results in the estimation of the variance of a characteristic for any probability distributions describing signals.

The above-listed limitations pose a barrier to normative applications of correlative relations in the final measurement assessments with stochastic relations. Therefore, in studies and publications, other probabilistic characteristics describing the relationships of a stochastic nature are introduced and applied [14]. Methods of measurement and analysis of such characteristics are being developed intensively.

In performing a metrological identification of signals and systems, developing models and making research into stochastic signals, the authors use the theory and techniques of conditional signal averaging. In the paper there is examined a real approximate model of exponential correlation and its impact on the process of conditional data averaging. Continuity of the derivative of autocorrelation function was evidenced for the examined model of signal correlation. This makes possible the theoretical research into the correlation of conditionally averaged implementations of the signal. The conditions of averaging aimed at reducing the variance of the conditional average value were determined. The theoretical model of exponential correlation was compared with the results of experimental research.

## 2. Models of autocorrelation function

Linear and exponential models of the *autocorrelation function* (ACF) are most frequently applied in descriptions of data autocorrelation. Linear and exponential autocorrelation functions $R_x(\tau)$ of an argument $\tau = 0$ have a common feature – a lack of continuity of their derivatives, which in many situations makes analysis and calculations difficult when processing the signals. Such functions are called non-differentiable and they are characterized by the infinite value of the derivative variance.

A binary synchronous signal of parameters $A$, $T$ and with an even distribution of moments of changes in the signal value (Fig. 1a) is characterized by a linear autocorrelation function (Fig. 1b). These kinds of signals occur in digital processing systems, *e.g.* after sampling and quantization; the analogue signals are transformed into bivalent signals.

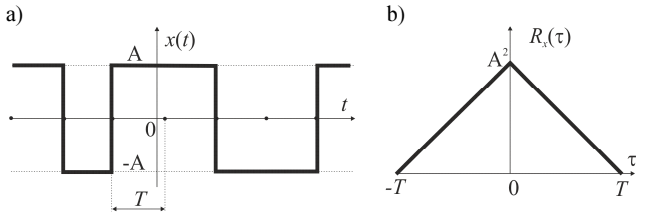



Fig. 1. A binary synchronous signal: a waveform (a); an autocorrelation function (b).

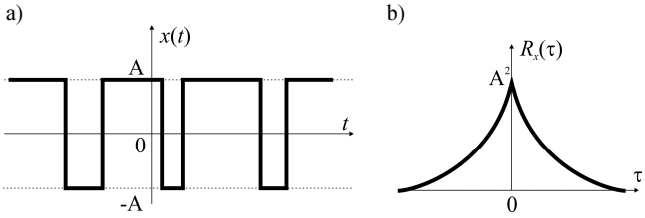



Fig. 2. A binary asynchronous signal: a waveform (a); an autocorrelation function (b).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 69–78.

Binary asynchronous signals with a random distribution of moments of changes in the signal value (Fig. 2a) are characterized by the ACF with an exponential shape, as presented in Fig. 2b. These kinds of signals occur in radioactive radiation trajectories.

An autocorrelation function with an exponential shape is also obtained by the output signal of a low-pass RC filter, when applying a white noise filter with a constant power spectral density $G_x(f) = G_0$ at the input:

$$R_x(\tau) = \frac{G_0}{RC} e^{-\frac{|\tau|}{RC}}. \tag{1}$$

The exponential shape of the ACF is a relatively frequent model when processing and describing analogue stochastic signals. In practice, an approximate model of exponential correlation is obtained for signals with a limited bandwidth with the features of white noise, passing physical inertial systems. Distributions of physical signals are usually normal or quasi normal, due to the central limit theorem and the inertia of typical processing systems.

## 3. Conditional averaging of auto-correlated data

The autocorrelation function is the main characteristic in the time domain describing the relation of stochastic signals. As a mixed second-order moment, it creates certain calculation difficulties, especially in assessment of the characteristic variance of auto-correlated data. The scope of usefulness of the ACF is narrowed down to linear models in probabilistic relations.

Restrictions for the measurement applications of correlation characteristics cause seeking other forms of description of stochastic relations for signals in the time domain. New possibilities in this scope for linear and non-linear models in metrological applications appear thanks to the use of functional and numerical conditional characteristics, in particular those of the conditional expected value and the conditional variance [14, 15]. The conditional expected value ensures the best estimate of interdependencies of stochastic signals in the mean square sense.

The conditional expected value in the time domain as the first-order central moment is described by relatively simple mathematical models with equally simple algorithms of their practical implementation corresponding to them [15]. In metrological applications of conditional averaging a right selection of the averaging condition enables to reduce the variance of estimates of experimental characteristics, which is one of the main objectives in measurement.

In basic applications of conditional averaging of Gaussian random signals, the characteristics of linear regression are used. For a single stationary signal with a distribution $N(0, \sigma_x)$ and a normalized ACF $\rho_x(\tau)$, the conditional expected value and the conditional variance are described by the following relations:

$$E(x_2|x_1) = \rho_x(\tau)x_1, \tag{2}$$

$$Var(x_2|x_1) = \sigma_x^2\left(1 - \rho_x^2(\tau)\right), \tag{3}$$

where: $x_1$ and $x_2$ are values of signal $x(t)$ at moments $t_1$ and $t_2$, respectively; $\tau = t_2 - t_1$.

In a simplified model of averaging non-correlated $M$ fragments of $x(t)$, after exceeding the $x(t) = x_p$ level [14], the assessment of the relative standard uncertainty of the conditional value of arithmetic mean is:

$$\varepsilon = \frac{\sigma_x}{\sqrt{M}x_p} \frac{\sqrt{1-\rho_x^2(\tau)}}{\rho_x(\tau)}. \tag{4}$$

In algorithms of conditional averaging using the maximum number of conditions $x(t) = x_p$ initiating the averaging, correlation of subsequently averaged fragments of the signal becomes problematic.

The following part of the paper presents the results of studies into correlation depending on a level $x_p$, which initiates conditional averaging. The studies were carried out assuming a normal distribution and an exponential correlation of a signal $x(t)$.

## 4. Assessment of auto-correlated data

Transition of white noise with a flat power spectral density in the $B$ band, equal to $G(\omega) = \sigma^2 / B$, through an $RC$ inertial system is described by the relation of one-sided spectral density at the system output:

$$G_x(\omega) = G(\omega)\left(\frac{1}{\sqrt{1+(\omega RC)^2}}\right)^2 = \frac{\sigma^2}{B\left[1+(\omega RC)^2\right]}. \tag{5}$$

The ACF at the inertial system output is described by:

$$R_x(\tau) = \frac{1}{2\pi}\int_0^{2\pi B} G_x(\omega)\cos\omega\tau d\omega = \frac{\sigma^2}{2\pi B}\int_0^{2\pi B}\frac{\cos\omega\tau}{1+(\omega RC)^2}d\omega\,. \tag{6}$$

After transformations, for $\tau = 0$ we arrive at:

$$R_x(0) = \sigma_x^2 = \frac{\sigma^2}{2\pi BRC} arctag\, 2\pi BRC\,. \tag{7}$$

**Example 1:**

$B = 25\cdot 10^3\ Hz$,

$RC = 100\cdot 10^{-6}\ s$,

$2\pi BRC = 2\pi\cdot 25\cdot 10^3\cdot 10^2\cdot 10^{-6} = 5\pi$,

$$R_x(0) = \sigma_x^2 = \frac{\sigma^2}{2\pi BRC}\frac{\pi}{2} = \frac{\sigma^2}{4BRC} = \frac{\sigma^2}{10}.$$

The autocorrelation of subsequent instances exceeding a given level $x_p$ by a signal $x(t)$ can be determined provided that the ACF $\rho_x(\tau)$ and statistical assessments of time intervals between appropriate instances exceeding the level $x_p$ are known.

In order to assess the autocorrelation of subsequent signal fragments, after exceeding the given level $x_p$, the ratio of the maximum interval correlation $\tau_{km}$ and the average interval $\bar{\tau}_p$ is determined for the signal $x(t)$ exceeding the level $x_p$.

The average time of passing the level $x_p$ by a signal $x(t)$ is described by [16]:

$$\bar{\tau}_p = \frac{1}{\bar{M}(x_p)} = \frac{2\pi}{\omega_{1x}} e^{\frac{x_p^2}{2\sigma}}, \tag{8}$$

where: $\bar{M}(x_p)$ $\tau$ an average number of signals $x(t)$ passing the level $x_p$ with a derivative of one sign in a given time unit; $\omega_{1x}$ – an average frequency of the spectrum of a random signal $x(t)$ determined by:

$$\omega_{1x}^2 = \sqrt{-\rho_x''(0)} = \frac{1}{2\pi\sigma_x^2}\int_0^{2\pi B}\omega^2 G_x(\omega)d\omega\,, \tag{9}$$

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 69–78.

where: $\rho_x''(0)$ – the second derivative of normalized ACF of a signal $x(t)$ for $\tau = 0$. Taking (5) and (9) into account, after necessary calculations, we obtain:

$$\omega_{1x} = \sqrt{-\rho_x''(0)} = \sqrt{\frac{1}{(RC)^2}\left(\frac{2\pi BRC}{arctg\,2\pi BRC} - 1\right)}. \tag{10}$$

Based on the relations (8) and (10), the average time between signals $x(t)$ passing the level $x_p$ with a derivative of one sign is:

$$\bar{\tau}_p = \frac{2\pi}{\sqrt{\frac{1}{(RC)^2}\left(\frac{2\pi BRC}{arctg\,2\pi BRC} - 1\right)}} e^{\frac{x_p^2}{2\sigma_x^2}}. \tag{11}$$

For the maximum interval of autocorrelation $\tau_{k_m}$ of a signal $x(t)$ with an exponential ACF, equal to $\tau_{k_m} = 3\tau_k = 3RC$, the ratio $\bar{\tau}_p / \tau_{k_m}$ is described by:

$$\frac{\bar{\tau}_p}{\tau_{k_m}} = \frac{2\pi}{\sqrt{\frac{1}{(RC)^2}\left(\frac{2\pi BRC}{arctg\,2\pi BRC} - 1\right)} \cdot 3RC} e^{\frac{x_p^2}{2\sigma_x^2}} = \frac{\frac{2}{3}\pi}{\sqrt{\frac{2\pi BRC}{arctg\,2\pi BRC} - 1}} e^{\frac{x_p^2}{2\sigma_x^2}}. \tag{12}$$

**Example 2:**

For given relative values of the level $\nu_p = x_p / \sigma_x$ initiating conditional averaging and for the data included in Example 1, the ratio $\bar{\tau}_p / \tau_{k_m}$ is:

$$\frac{\bar{\tau}_p}{\tau_{k_m}} = \frac{\frac{2}{3}\pi}{\sqrt{\frac{5\pi}{\frac{\pi}{2}} - 1}} e^{\frac{x_p^2}{2\sigma_x^2}} = \frac{2}{9}\pi e^{-\frac{\nu_p^2}{2}}. \tag{13}$$

Calculated and rounded values of the ratio $\bar{\tau}_p / \tau_{k_m}$ for several values $\nu_p$ are presented in Table 1.

Table 1. The values of ratio $\bar{\tau}_p / \tau_{k_m}$.

| $\nu_p$ | 0 | 1 | $\sqrt{2}$ | 2 |
|---|---|---|---|---|
| $\frac{\bar{\tau}_p}{\tau_{k_m}}$ | 0.70 | 1.15 | 1.90 | 5.16 |

For values $x_p \geq \sigma_x$, the averaged implementations of a signal $x(t)$ exceeding the level $x_p$ with a derivative of one sign can be practically considered to be non-correlated.

Average implementations of $x(t)$, initiated by subsequent instances exceeding the level $x_p$ with a derivative of any sign, can be described by the average time of a signal $x(t)$ being above the level $\nu_p = x_p / \sigma_x$:

$$\bar{\tau}_{p\pm} = \frac{\pi}{\omega_{1x}} e^{\frac{x_p^2}{2\sigma_x^2}} \left[1 - 2\Phi(\nu_p)\right], \tag{14}$$

where: $\Phi(\nu)=\frac{1}{\sqrt{2\pi}}\int_0^{\nu_p} e^{\frac{-z^2}{2}}\,dz$ – the Laplace's integral.

A chart of relation (14) is presented in Fig. 3.

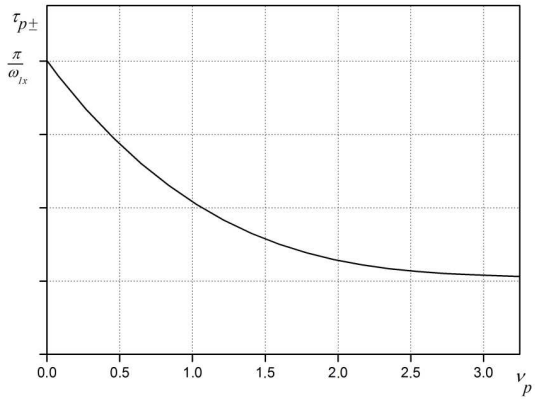



Fig. 3. An average time during which a signal $x(t)$ remains above the level $\nu_p$.

Calculated and rounded values of the ratio $\bar{\tau}_{p\pm}/\tau_{k_m}$ for several values $\nu_p$ are presented in Table 2.

Table 2. The values of ratio $\bar{\tau}_{p\pm}/\tau_{k_m}$.

| $\nu_p$ | 0 | 1 | $\sqrt{2}$ | 2 |
|---|---|---|---|---|
| $\frac{\bar{\tau}_{p\pm}}{\tau_{k_m}}$ | 0.35 | 0.18 | 0.15 | 0.12 |

It can be concluded from the provided comparison that the time intervals between subsequent instances exceeding the level $x_p \geq 0$ by a signal $x(t)$ are on average significantly lower than the maximum interval of correlation $\tau_{k_m}$. The ratio of the arithmetic mean of two time intervals between subsequent instances exceeding the level $\nu_p=\sqrt{2}$ with derivatives of various signs and the maximum correlation interval is:

$$\frac{\frac{\bar{\tau}_{p\pm}}{2}}{\tau_{k_m}}=\frac{\pi}{9}e^1\approx 0.95. \tag{15}$$

In the publications [17, 18] it was indicated that the optimal value of the level initiating conditional averaging of a signal with a normal distribution and an exponential ACF is included in the following interval:

$$\sqrt{2}\sigma_x \leq x_p \leq 2\sigma_x. \tag{16}$$

Using the obtained results to optimize the process of estimating the conditional expected value, we can perform the following sequence of calculations, basing on a random signal digitally registered in time:

1. Assume an averaging level $x_p$, $\sqrt{2}\sigma_x \leq x_p \leq 2\sigma_x$.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 69–78.

2. In the time interval $0-T_r$ $\left(T_r \geq \tau_{k_m}\right)$ average subsequent $M/2$ of implementations exceed the level $x_p$ with a positive derivative.
3. Start and perform averaging with a delay by the first implementation in time $T_r$ from the previous point also in the time interval $0-T_r$; subsequent $M/2$ of implementations exceed the level $x_p$ with a negative derivative.
4. Perform synchronic averaging of values of partial characteristics from points 2 and 3.

Assessment of the relative standard uncertainty of determining the conditional average value for $\left(0 \leq \tau \leq T_r\right)$ can be calculated from the relation (4).

## 5. Experimental studies

The low-pass white noise $x(t)$ with a distribution N(0 V, 0.3 V) and a frequency band $B = 25$ kHz was applied on the inputs of first-order inertial systems with three different time constants $T_c$ of: 10 µs, 30 µs and 100 µs. Figs. 4b–4d provide the obtained functions of the conditional average value (CAV) $\bar{x}(\tau)|x_p$, which are proportional to appropriate autocorrelation functions $\rho_x(\tau)$. For comparison, in Figure 4a the behaviour $\bar{x}(\tau)|x_p = f(\tau)$ of the original signal $x(t)$ was presented.

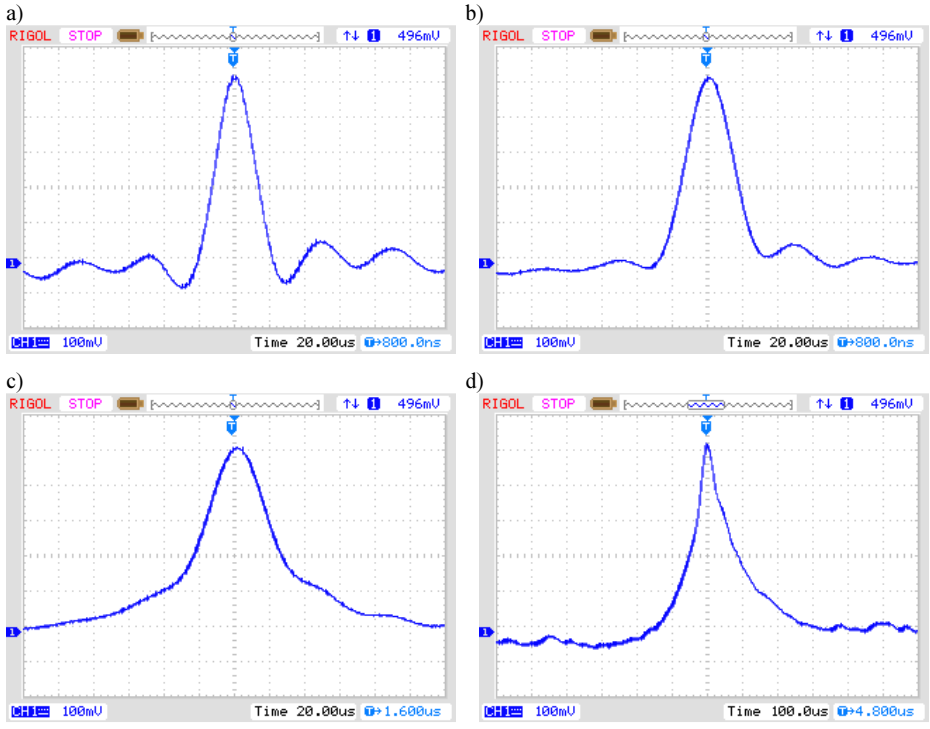



Fig. 4. The behaviour of functions of CAV: the low-pass white noise: N(0 V, 0,3 V), $B = 25$ kHz and after the noise passed the first-order inertial system with various time constants $T_c$ (a): $T_c = 10$ µs (b); $T_c = 30$ µs (c); $T_c = 100$ µs (d).

The characteristics were designated using a RIGOL digital oscilloscope, with the level initiating averaging of $x_p = 0.5$ V and the number of averages $M = 256$.

The compliance of the experimentally obtained values $\rho_x(k\Delta t)$, determined based on the conditional function of the average value, was confirmed by calculating the normalized value of autocorrelation from the relation $\rho(k)=e^{-k}$. The obtained analytical results for $T_c = 100$ µs, $\Delta t = 100$ µs and $k = \Delta t/T_c$ are presented in Table 3.

Table 3. The values of function $\rho(k)=e^{-k}$ obtained from calculations.

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $\rho(k)$ | 1 | 0.36 | 0.14 | 0.05 | 0.018 | 0.007 |

Figure 5 illustrates the behaviour of experimentally determined normalized ACF of the low-pass white noise: N(0 V, 0.3 V), $B = 25$ kHz when it passed the first-order inertial system with a time constant $T_c = 100$ µs. The values $\rho(k)$ presented in Table 3 were marked on the chart with crosses. A significant compliance of the results obtained experimentally and analytically is visible.

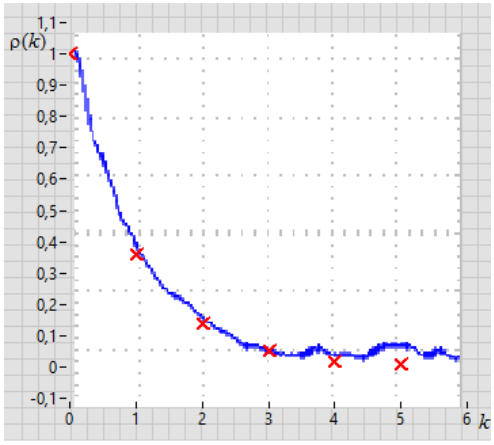



Fig. 5. The experimental behaviour of normalized autocorrelation function $\rho_x(k\Delta t)$ of the low-pass white noise N(0 V, 0.3 V), $B = 25$ kHz after it passed the first-order inertial system with a time constant $T_c = 100$ µs (continuous line) and the calculated autocorrelation values (points x).

## 6. Summary

1. In practice, the exponential autocorrelation model is obtained using white noise signals with a limited $B$ band, passing physical inertial systems with a time constant $T_c$. In the implemented experiment for a product $BT_c \geq 2.5$, the function $\rho_x(k\Delta t)$ for $\tau = 0$ has a derivative, and for $\tau > 0$ the function shape is exponential. The presented realistic signal model is useful in practical applications.
2. Due to simple theoretical and practical models, when developing the auto-correlated measurement data with normal and quasi-normal distributions, conditional averaging algorithms can be used. It is especially beneficial in the case of strong data autocorrelation. The conditional average value is proportional to the ACF, therefore it can be used to assess interdependencies of measurement data, *e.g.* exponentially auto-correlated data.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 69–78.

3. In exponential autocorrelation models for a threshold condition value $x_p \geq \sigma_x$, initiating subsequent conditional averaging, averaged implementations of a signal $x(t)$ exceeding the level $x_p$ with a derivative of one sign can be practically considered to be non-correlated.
4. In the exponential autocorrelation model, the subsequent averaged implementations of a signal $x(t)$ exceeding the level $x_p$ with a derivative of any sign are significantly correlated.
5. The process of conditional averaging a signal $x(t)$ can be optimized: through selecting a value of the level $x_p$, and average conditional components determined with a lack of data correlation for instances exceeding the level $|x_p|$ with derivatives of any sign.
6. When estimating the conditional expected value for exponential oscillatory data autocorrelation models, relatively large values of maximum intervals of correlation $\tau_{km}$ should be taken into account when averaging with the use of non-correlated samples with a time $T_p \geq \tau_{km}$. Assessing values and signs of autocorrelations when averaging using correlated samples also needs to be considered.

## References


[1] Guide to the expression of uncertainly in measurement. (1995). International Organisation for Standardisation.

[2] International vocabulary of basic and general terms in metrology (VIM). (2004). International Organization for Standardization. (Revision of the 1993 edition).

[3] Evaluation of measurement data. Supplement 1 to the "Guide to the expression of uncertainty in measurement" – Propagation of distributions using a Monte Carlo method (2008). JCGM 101:2008.

[4] Evaluation of measurement data. Supplement 2 to the "Guide to the expression of uncertainty in measurement" – Extension to any number of output quantities. (2011). JCGM 102:2011.

[5] Box, G.E.P., Jenkins, G.M., Reinsel, G.C. (1994). *Time series analysis: forecasting and control*. Prentice Hall, Englewood Cliffs.

[6] Bayley, G.V., Hammersley, G.M. (1946). The "effective" number of independent observations in an autocorrelated time-series. *J. Roy. Stat. Soc. Suppl.* 8, 184–197.

[7] Zhang, N.F. (2006). Calculation of the uncertainty of the mean of autocorrelated measurements. *Metrologia* 43, 276–281.

[8] Dorozhovets, M., Warsza, Z. (2007). Evaluation of the uncertainty type A of autocorrelated measurement observations. *Measurement Automation and Monitoring*, 53(2), 20–24.

[9] Witt, T.J. (2007). Using the autocorrelation function to characterize time series of voltage measurements. *Metrologia*, 44, 201–209.

[10] Zięba, A. (2010). Effective number of observations and unbiased estimators of variance for autocorrelated data – an overview. *Metrol. Meas. Syst.*, 17(1), 3–16.

[11] Zięba, A., Ramza P. (2011). Standard deviation of the mean of autocorrelated observations estimated with the use of the autocorrelation function estimated from the data. *Metrol. Meas. Syst.*, 18(4), 529–542.

[12] Dorozhovets, M. (2009). Influence of lack of a priori knowledge about autocorrelation functions of observations on estimation of their average value standard uncertainty. *Measurement Automation and Monitoring*, 55(12), 989–992.

[13] Zhang, N.F. (2008). Allan variance of time series models for measurement data. *Metrologia*, 45, 549–561.

[14] Kowalczyk, A., Szlachta, A., Hanus, R. (2012). Standard uncertainty determination of the mean for correlated data using conditional averaging. *Metrol. Meas. Syst.*, 19(4), 787–796.

[15] Kowalczyk, A. (2015). *Measurement applications of conditional signal averaging.* Oficyna Wydawnicza Politechniki Rzeszowskiej, Rzeszów.

[16] Bendat, J.S., Piersol, A.G. (2010). *Random data. Analysis and measurement procedures.* John Wiley & Sons, New York.

[17] Kowalczyk, A. (2008). Classical method for determination of linear system dynamic properties using signal conditional averaging. *Measurement Automation and Monitoring*, 54(12), 820–823.

[18] Szlachta, A., Kowalczyk, A., Wilk, G. (2009). Accuracy investigations of impulse response estimation obtained by conditional averaging. *Measurement Automation and Monitoring*, 55(12), 981–984.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 79–89.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# PROGRAMMABLE INPUT MODE INSTRUMENTATION AMPLIFIER USING MULTIPLE OUTPUT CURRENT CONVEYORS

**Bogdan Pankiewicz**

*Gdańsk University of Technology, Faculty of Electronics, Telecommunication and Informatics, G. Narutowicza 11/12, 8-233 Gdańsk, Poland (✉ bpa@eti.pg.gda.pl, +48 58 347 1974)*

**Abstract**

In this paper a programmable input mode *instrumentation amplifier* (IA) utilising second generation, multiple output current conveyors and transmission gates is presented. Its main advantage is the ability to choose a voltage or current mode of inputs by setting the voltage of two configuration nodes. The presented IA is prepared as an integrated circuit block to be used alone or as a sub-block in a microcontroller or in a *field programmable gate array* (FPGA), which shall condition analogue signals to be next converted by an *analogue-to-digital converter* (ADC). IA is designed in AMS 0.35 μm CMOS technology and the power supply is 3.3 V; the power consumption is approximately 9.1 mW. A linear input range in the voltage mode reaches ± 1.68 V or ± 250 μA in current mode. A passband of the IA is above 11 MHz. The amplifier works in class A, so its current supply is almost constant and does not cause noise disturbing nearby working precision analogue circuits.

Keywords: instrumentation amplifier, current conveyor, programmable analogue circuit.



## 1. Introduction

*Instrumentation amplifiers* (IA) are usually employed as input stages in a variety of applications. Their main purpose is to amplify desired differential signals while simultaneously suppressing the unwanted common mode ones. The application area covers sensor signal amplification, medical instrumentation, data acquisition and many others [1–5]. Most IAs work in the voltage mode (*i.e.* input and output signals are voltage ones) and are built using *operational amplifiers* (OA) and resistor networks [3–5]. New integrated circuit manufacturing technologies enable to use smaller power supply voltages, which resulted in turning to signal processing in the form of current instead of voltage and creation of new active blocks called current conveyors [6–9]. As a result many IAs are built of current conveyors and IAs working in all four possible modes are also in common use [1–2]. In this paper a new concept of *multiple output second order current conveyor* (MOCCII) is presented (Section 2). It uses, as the output stage, a current amplifier reported in [11]. Then MOCCII is used as the main active block of a programmable input mode instrumentation amplifier (Sections 3 and 4). There are numerous IAs with a programmable gain and offset known in the literature [3–5] but there is a lack of a programmable input mode IA. The presented, programmable input mode IA is designed as an integrated circuit block to be used alone or as a sub-block of a microcontroller or FPGA, which can condition analogue signals to be next converted by a system *analogue to digital converter* (ADC).

## 2. Multiple output current conveyor circuit

Since their introduction [6], current conveyors have been widely used in analogue signal processing applications. To date, many variations of current conveyors have been presented,

Article history: received on Nov. 16, 2016; accepted on Dec. 30, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0017.

both with positive and negative current gains, their generations being marked from I to III, and also having multiple outputs [7–10]. In this paper, a *multiple output second-generation current conveyor* (MOCCII) is presented and used as an active block of IA. It has inputs Y and X and 3 outputs: $Z_P$, $Z_{M1}$ and $Z_{M2}$. Output $Z_P$ is the positive one while $Z_{M1}$ and $Z_{M2}$ are two independent negative outputs. The MOCCII graphical symbol and its terminals' voltages and currents are defined in Fig. 1. A matrix equation describing ideal electrical properties of an MOCCII element is given by:

$$\begin{bmatrix} i_Y \\ v_X \\ i_{ZP} \\ i_{ZM1} \\ i_{ZM2} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_Y \\ i_X \\ i_{ZP} \\ i_{ZM1} \\ i_{ZM2} \end{bmatrix}. \tag{1}$$

From (1) it is apparent that an MOCCII element is equivalent to a second-generation current conveyor (CCII) having 2 negative and 1 positive independent outputs. The main advantage of the below shown structure over the previously presented one [7–9] is simultaneous providing 3 independent outputs, each of them exhibiting similar and good frequency responses. Such properties are usually impossible to obtain using the current mirror cascading technique for generating the negative output.


Fig. 1. A graphic symbol of the proposed *multiple output second-generation current conveyor* (MOCCII) block.

### *2.1. Architecture of MOCCII*

The architecture of proposed MOCCII is presented in Fig. 2. It consists of an *operational amplifier* (OA) and – as the output stage – a current amplifier described in [11]. Due to the negative feedback loop in the signal path (which starts at X terminal and then passes through OA and an inverting MN1 device), voltages at Y and X terminals should be of the same value. In a real circuit the equality of voltages at Y and X terminals depends mainly on mismatches of devices in the input stage of OA, and thus this sub-circuit should be designed carefully using relatively large devices and a high overdrive voltage [12]. Simultaneously, any current going via X terminal is the input current flowing to the current amplifier stage, which is marked in Fig. 2 with a dashed line. That is why any current flowing to the X terminal is also amplified and moved to output terminals $Z_P$, $Z_{PM1}$, and $Z_{M2}$. It should be noted that – according to the current amplifier concept in [11] – if all $M_{N1} - M_{N4}$ devices have identical dimensions, then, ideally, neglecting resistances of bias current sources $I_{BIAS}$ and $4I_{BIAS}$, the current gains to output terminals $Z_P$, $Z_{PM1}$, and $Z_{M2}$ are equal to –1, 1 and 1, respectively. In a real circuit, which employs cascoded MOS current sources, the absolute values of those current gains are very close but not exactly equal to 1. It is also worth noticing that, according to the circuit in [11], the current gains to all outputs have similar frequency responses.

The output stage of MOCCII works in class A. This implies a constant value of current supply and also limits output current signals to a range of $\pm I_{BIAS}$. The current forced by X terminal is also limited to a range $\pm I_{BIAS}$.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 79–89.

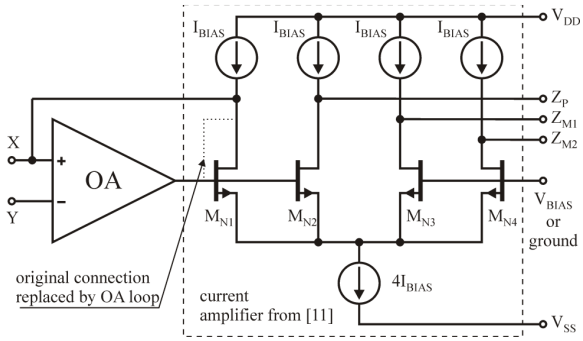



Fig. 2. A general architecture of the multiple output second-generation current conveyor. Instead of a diode connection of the device MN1 (dotted line), an *operational amplifier* (OA) is employed in the negative feedback loop. A dashed line surrounds the employed current amplifier presented in [11].

### *2.2. Implementation of MOCCII*

A design of MOCCII using the architecture from Fig. 2 is presented in Fig. 3. The OA is built using a single pMOS differential pair (devices $M_{OA1}$–$M_{OA2}$) with nMOS current mirror as load (devices $M_{OA3}$ – $M_{OA4}$). A transistor $M_{OA5}$ is the current source for the input differential pair and its dimensions were chosen to obtain a current of approximately 50 μA. Devices $M_{N1}$–$M_{N4}$ constitute the core of the current amplifier. A bias current source $4I_{BIAS}$ is built with the use of a low voltage cascode current source created by devices $M_{NB1}$–$M_{NB4}$. The diode connected devices $M_{NB5}$, $M_{PB11}$ and $M_{PB12}$ generate a constant voltage feed to the gates of $M_{NB1}$ and $M_{NB3}$, which is necessary for the proper operation of the low voltage cascode. Similarly, devices $M_{PB1}$–$M_{PB10}$, $M_{PB13}$, $M_{NB6}$ and $M_{NB7}$ create four low voltage cascode current sources $I_{BIAS}$. The circuit is biased using a constant current of value $I_{BIAS}$ = 250 μA. The obtained value of voltage gain measured from Y to X terminal exceeds 10k V/V, which is sufficiently high to obtain a near unity voltage gain seen from Y to X terminals when the negative feedback loop is closed through connection of X and $X_{FB}$ terminals.

This feedback loop connection is accomplished outside the MOCII circuit in the IA structure using transmission gates. Separation of X and $X_{FB}$ terminals enables to sense a voltage across the resistance $R_X$, without an extra nonlinear component formed on the transmission gate, which is used as a programmable connection. The $C_{COMP}$ capacitor is necessary to maintain stability of the negative feedback loop.

The MOCCII circuit was designed and simulated in a Cadence Virtuoso IC6.1.5 environment using an AMS 350nm CMOS technology kit. The circuit test setup is presented in Fig. 4. Simulations were carried out with the connection of Y terminal to a voltage signal source, while X and $X_{FB}$ terminals were tied together and connected via a 7kΩ resistor to the signal ground. Output currents were measured at circuit shorts of $Z_P$, $Z_{M1}$ and $Z_{M2}$ terminals to the signal ground. The simulated results are given in Table 1 and in Fig. 5.

The MOS devices' dimensions given in Fig. 3 are quite high in comparison with the technological minimal ones. The dimensions of transistors were chosen with careful consideration of process mismatches. According to [12] matching of device parameters is reversely proportional to the square root of the device area. In order to obtain an acceptable level of current and voltage offsets as well as acceptable changes of current gain of the conveyor, the design process was started with almost minimal technological dimensions and then the device dimensions were gradually increased until satisfactory parameter values were obtained. If the presented values are not satisfactory in a certain application, a further increase of devices' dimensions can be made, but the square root relationship should be borne in mind,

regarding that better mismatch parameters cost a large circuit area. A big area of devices causes also higher capacitances of devices and – as a result – worse frequency responses are obtained. Especially, increasing a length of MOS transistors implies a significant bandwidth loss of approximately quadratic dependence, so widths of transistors should be expanded firstly and their lengths – only if necessary.

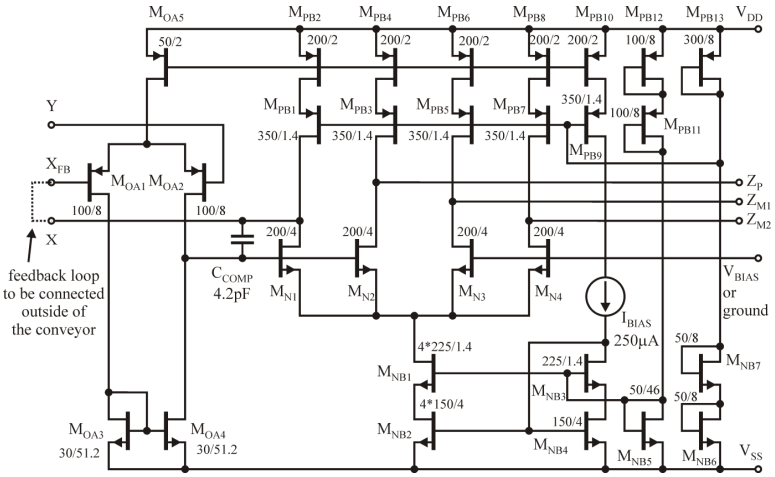



Fig. 3. A detailed scheme of MOCCII designed in AMS 350nm CMOS technology; transistor dimensions are in µm; format is as follows: optional multiplier*width/length.

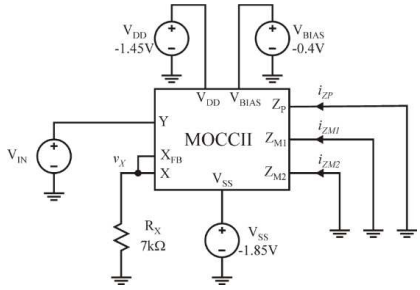



Fig. 4. The testing environment for the amplifier from Fig. 3.

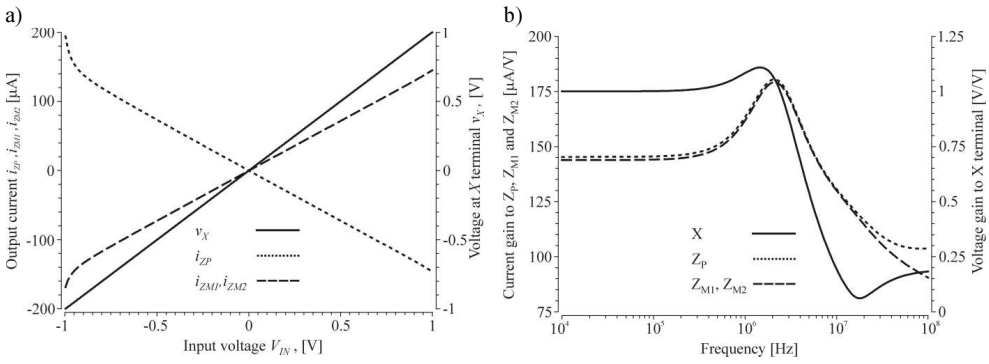



Fig. 5. The simulated characteristics of the MOCCII from Fig. 3 obtained using the testing environment presented in Fig. 4. DC transfer characteristics (a); small signal frequency responses (b).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 79–89.

Table 1. The simulated parameters of the MOCCII from Fig. 3 obtained with using the testing environment presented in Fig. 4.

| Parameter name | Unit | Value |
|---|---|---|
| Power supply voltage $V_{DD}$ – $V_{SS}$ | V | 3.3 |
| Current consumption | uA | 1380 |
| Voltage gain from Y to X | V/V | 0.9998 |
| 3dB passband of Y to X gain | MHz | 3.91 |
| Standard deviation of voltage gain from Y to X, result of 200 runs of MC analysis | μ V/V | 45.6 |
| Transconductance from Y to $Z_P$ @ Rx = 7 kΩ | μS | 145.5 |
| Frequency of 1deg phase loss at $Z_P$ output | MHz | 1.16 |
| Standard deviation of transconductance to $Z_P$, result of 200 runs of MC analysis | nS | 188.7 |
| Transconductance from Y to $Z_{M1}$ and $Z_{M1}$ @ Rx = 7 kΩ | μS | 144.1 |
| Frequency of 1deg phase loss at $Z_{M1}$ and $Z_{M2}$ outputs | MHz | 1.10 |
| Standard deviation of transconductance to $Z_{M1}$ and $Z_{M2}$, result of 200 runs of MC analysis | nS | 262.5 |
| Resistance of X and $X_{FB}$ terminals tied together for low frequencies | Ω | 1.35 |
| Resistance of $Z_P$ terminal for low frequencies | kΩ | 371 |
| Resistance of $Z_{M1}$and $Z_{M2}$ terminals for low frequencies | kΩ | 722 |
| Equivalent Y input capacitance | fF | 790 |
| Equivalent $Z_P$, $Z_{M1}$ and $Z_{M2}$ capacitance | fF | 563 |
| Noise at $Z_P$ output @100 kHz | $pA/\sqrt{Hz}$ | 10.2 |
| Noise at $Z_{PM1}$ and $Z_{PM2}$ outputs @100 kHz | $pA/\sqrt{Hz}$ | 16.9 |
| Offset voltage at X terminal for $V_Y$ = 0 V | μV | 181.1 |
| Standard deviation of input referred offset voltage, result of 200 runs of MC analysis | μV | 583 |
| Offset output current at $Z_P$ for $V_Y$ = 0 V | nA | 0.489 |
| Standard deviation of $Z_P$ output current offset, result of 200 runs of MC analysis | nA | 1335 |
| Offset output currents at $Z_{PM1}$ and $Z_{PM2}$ for $V_Y$ = 0 V | nA | 199.3 |
| Standard deviation of $Z_{PM1}$ and $Z_{PM2}$ outputs current offset, result of 200 runs of MC analysis | nA | 2219 |

## 3. Instrumentation amplifier design

A circuit diagram of the programmable input mode instrumentation amplifier is presented in Fig. 6. It consists of 2 MOCCII conveyors from Fig. 3 and 16 transmission gates. A schematic of the transmission gate is presented in the right bottom part of Fig. 6. It consists of nMOS and pMOS transistors and an inverter gate. The inverter gate is used only for generating the opposite logic level for the pMOS device. If a supply voltage $V_{DD}$ is applied at the control input EN, then both transistors are working in the ohmic region. The gate is in the closed state and an equivalent resistance occurs between $IO_1$ and $IO_2$ terminals with a value of approximately a few hundred ohms . This resistance is slightly nonlinear but due to splitting X and $X_{FB}$ terminals in MOCCII the nonlinearity may be eliminated in the final IA. If EN input is supplied with a $V_{SS}$ voltage, then both MOS devices are cut off and the gate is in the open state. The resistance between $IO_1$ and $IO_2$ terminals is very high, theoretically infinitive. The dimensions of transistors in the transmission gate are small in comparison with the MOCCII circuit size and the use of 16 such items does not increase the overall area of IA by much. Feeding the control input of transmission gates with a $V_{DD}$ or $V_{SS}$ voltage constitutes the

final circuit configuration and the resulting mode of work. If the terminal VM is fed with a $V_{DD}$ voltage, then the amplifier works in the input voltage mode. Analogously, if the terminal CM is fed with a $V_{DD}$ voltage, then IA from Fig. 6 works in the input current mode. Equivalent circuit schematics for the voltage and current modes are presented in Fig. 7. The gates being in the open state are omitted, while the gates in the closed state are treated as short circuits with the exception of gates transferring a high current flowing to X terminal of MOCCII, which are represented by a resistance $R_G$. Both circuits have independent positive and negative current mode outputs $OUT_P$ and $OUT_M$.

Let the input stage of the MOCCII from Fig. 3 be considered once again, working without connecting together X and $X_{FB}$ terminals. If such a situation occurs, the X terminal may be treated as the output of transconductance amplifier built of a differential pair $M_{OA1}$, $M_{OA2}$ with a current mirror $M_{OA3}$, $M_{OA4}$ load and the second stage built of a transistor $M_{N1}$ in the CS configuration, whose transconductance decreases twice due to working in an extended differential pair $M_{N1}$–$M_{N4}$. Thus, the approximate, small signal transconductance value in the path from the input differential pair to the current flowing out of X terminal may be expressed as:

$$gm_X = \frac{i_X}{v_Y - v_{XFB}} = -gm_{MOA1}\left(r_{MOA2} \parallel r_{MOA4}\right) \cdot \frac{1}{2} gm_{MN1}, \tag{2}$$

where: $i_X$ is a current flowing out of X terminal; $v_Y$ and $v_{XFB}$ are respective voltages at Y and $X_{FB}$ terminals; $gm_{MOA1}$ is a small signal transconductance of device $M_{OA2}$; $r_{MOA2}$ and $r_{MOA4}$ are output resistances of devices $M_{OA2}$ and $M_{OA4}$, respectively; and $gm_{MN1}$ is a small signal transconductance of transistor $M_{N1}$. The value of (2) is quite high and a simulated value for the circuit in Fig. 3 is equal to 0.738 S. The node $X_{FB,}$ which senses the feedback signal, is placed directly at $R_X$ resistor and thus any resistor $R_G$ representing the transmission gate does not change the transfer characteristic of IA. Please notice that in the above circuit the current flowing out of X terminal flows also out of $Z_P$ and into $Z_{M1}$ and $Z_{M2}$ terminals. Due to this, output currents for the input voltage mode of IA presented in Fig. 7a may be expressed as:

$$i_{OUTP} = -i_{OUTM} = 3\frac{1}{\frac{2}{gm_X} + R_X} V_{IN}\Big|_{R_X >> 2/gm_X} \approx \frac{3}{R_X} V_{IN}, \tag{3}$$

where: $R_X$ is a resistance connected between $X_{FB}$ terminals of MOCCII conveyors and $gm_X$ is a small signal transconductance given by (2).

IA working in the current mode has both Y terminals connected together (to GND_CM node) and MOCCII amplifiers – due to the negative feedback loop – try to maintain the same voltage at $X_{FB}$ terminals. It works like a transconductance amplifier with its input and output tied together. The equivalent input resistance of a single node in respect to GND_CM is thus equal to:

$$\frac{1}{2} R_{IN\_CM} = \frac{1}{gm_X}, \tag{4}$$

and the differential mode resistance is twice as high. The current flowing into X terminal also flows to $Z_P$ and out of $Z_{M1}$ and $Z_{M2}$ terminals, thus output currents may be expressed by:

$$i_{OUTP} = -i_{OUTM} = -2I_{IN}. \tag{5}$$

Please notice that for the current mode only 2 outputs are tied together. Such a connection was chosen to obtain a high CMRR factor. The common mode input currents, due to summation at the outputs, are automatically eliminated only if current gains to $Z_P$ and $Z_{M1}$ are of exactly opposite values.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 79–89.

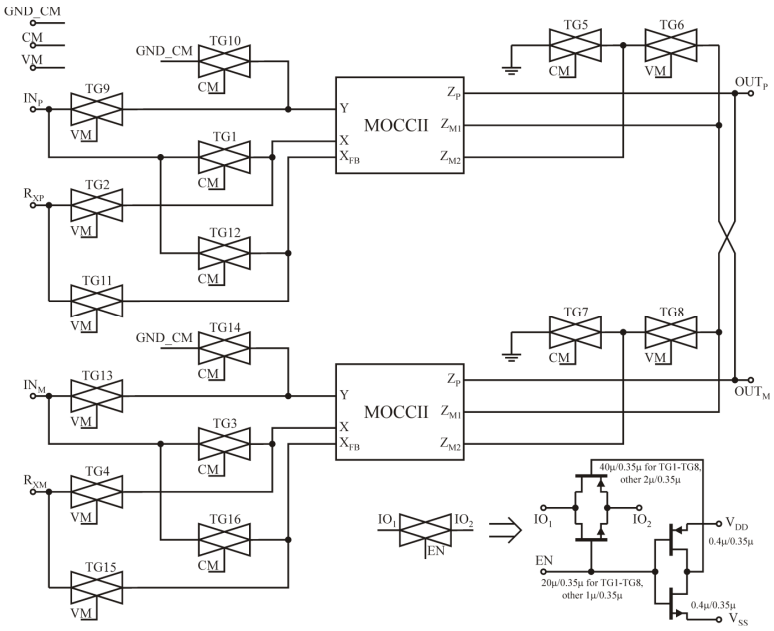



Fig. 6. A programmable input mode instrumentation amplifier using MOCCIIs and transmission gates.

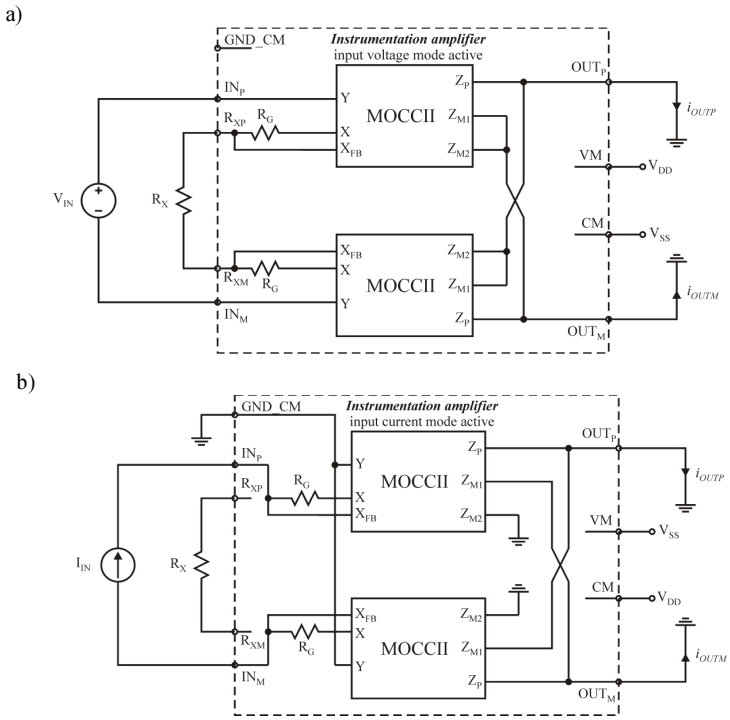



Fig. 7. Equivalent IA schematics in the input voltage (a) and the input current (b) modes together with the desired signal sources and a resistance $R_X$ necessary for proper operation only in the voltage mode. Output currents are measured at shorts to the signal ground. $R_G$ represents resistances of the transmission gate connected to X terminals of MOCCIIs.

## 4. Simulation results of instrumentation amplifier

The instrumentation amplifier from Fig. 6 working both in the input voltage and current modes has been simulated in detail. Testing environments such as in Fig. 7 with $R_X = 14$ kΩ and additional common mode input sources for simulation of CMRR factor were used. The simulation results are presented in Figs. 8–11 and in Tables 2 and 3. The power consumption is the same for both modes and is presented only in Table 2.

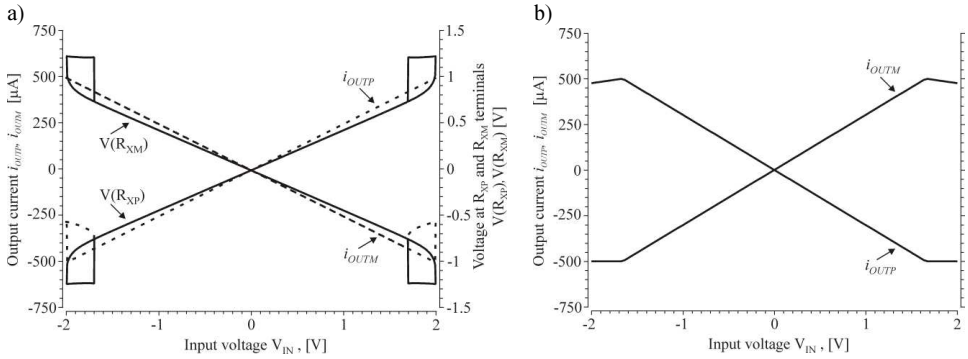



Fig. 8. DC transfer responses of the amplifier from Fig. 6 in (a) the voltage and (b) current modes.

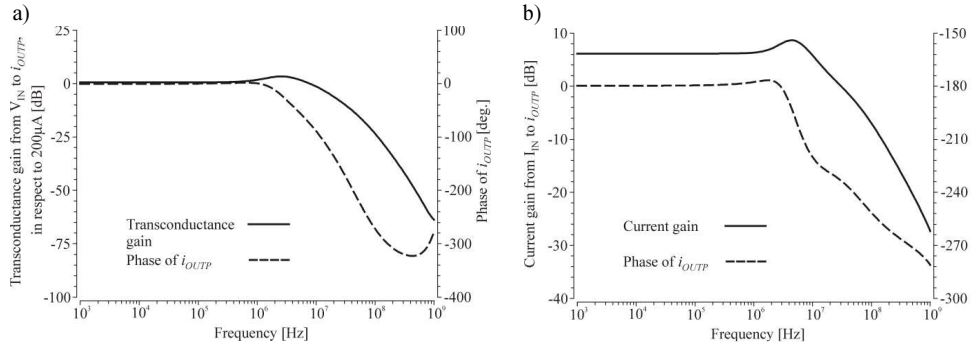



Fig. 9. AC responses of the amplifier from Fig. 6 in (a) the voltage and (b) current modes.

The DC transfer characteristics are presented in Fig. 8. In voltage mode, for the lowest and highest input voltages, a hysteresis is observed. This is caused by a limited operational range of the input MOS pair and to avoid entering this region by the circuit the input level should be limited. It is not observed in the current mode because the reference input GND_CM is connected to the signal ground. The simulated worst-case values of CMRR for 200 runs of MC simulation are equal to 55.3 dB and to 51.4 dB for the voltage and current modes, respectively.

Small signal frequency responses are presented in Fig. 9. Both configurations have a 3 dB frequency located over 11 MHz. Histograms of gains at 100Hz are presented in Fig. 11. Due to relatively large transistors used in MOCCII blocks, standard deviations of gains are small and their values are equal to about 0.13% of the desired gain.

The result of transient analysis for a 10 kHz harmonic signal in the form of THD factor is presented in Tables 2 and 3. Fig. 11 presents the results of transient analysis for a square wave of 100 kHz frequency. There are visible overshoots in the output signal. The current mode IA has a higher passband and higher slope rates at the output signal. A dynamic range defined as a ratio of the RMS value of input signal causing THD = 1% and the input referred integrated

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 79–89.

noise in a 100 Hz –1 MHz passband is equal to 77.3 dB and 85.1 dB for the input voltage and current modes, respectively.

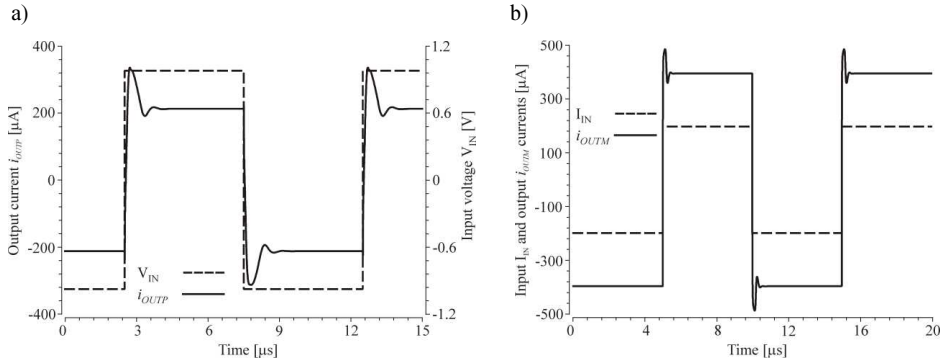



Fig. 10. Transient responses of the amplifier from Fig. 6 in (a) the voltage and (b) current modes. As the input pulse voltage or current sources of 100 kHz frequency were used, respectively.

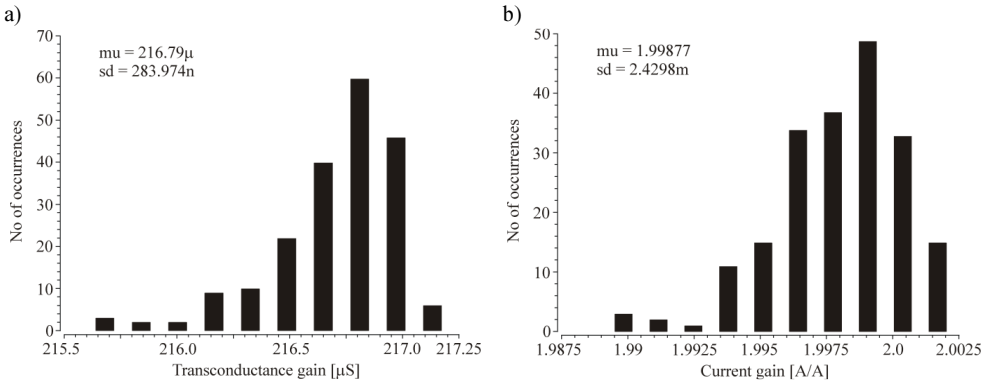



Fig. 11. Histograms of gains for the amplifier from Fig. 6 in (a) the voltage and (b) current modes. The result of 200 MC simulation runs with both mismatch and process changes.

Table 2. The simulated parameters of the instrumentation amplifier from Fig. 6 working in the voltage mode according to Fig. 7a.

| Parameter name | Unit | Value |
|---|---|---|
| Power supply voltage $V_{DD}$ – $V_{SS}$ | V | 3.3 |
| Current consumption | µA | 2763 |
| Transconductance gain for Rx = 14 kΩ | µA/V | 216.9 |
| Standard deviation of transconductance, result of 200 runs of MC analysis | µA/V | 0.284 |
| 3dB passband of transconductance gain | MHz | 11.95 |
| Output current range | µA | ±750 |
| Input referred noise spectral density @100 kHz | $nV/\sqrt{Hz}$ | 160.3 |
| Input referred integrated noise in bandwidth 100 Hz – 1 MHz | µV | 152.3 |
| Amplitude of 10 kHz harmonic signal for output current THD =–40dB | mV | 1670 |
| Dynamic range | dB | 77.3 |
| Worst case CMRR for 200 MC runs | dB | 55.3 |
| Output resistance | kΩ | 352.5 |
| Equivalent output capacitance | fF | 2005 |
| Differential input capacitance | fF | 412.3 |
| Input referred offset voltage | mV | 1.82 |
| Standard deviation of input referred offset voltage, result of 200 runs of MC analysis | mV | 18.7 |

Table 3. The simulated parameters of the instrumentation amplifier from Fig. 6 working in the current mode according to Fig. 7b.

| Parameter name | Unit | Value |
|---|---|---|
| Current gain | A/A | –2 |
| 3dB passband of current gain | MHz | 16.28 |
| Standard deviation of current gain, result of 200 runs of MC analysis | mA/A | 2.43 |
| Input referred noise spectral density @100k Hz | $pA/\sqrt{Hz}$ | 9.899 |
| Input referred integrated noise in bandwidth 100 Hz – 1M Hz | nA | 9.778 |
| Amplitude of 10kHz harmonic signal for output current THD = –40 dB | μA | 248 |
| Dynamic range | dB | 85.1 |
| Worst case CMRR for 200 MC runs | dB | 51.4 |
| Input differential resistance for low frequencies | Ω | 2.70 |
| Input referred offset current | nA | 98.9 |
| Standard deviation of input referred offset current, result of 200 runs of MC analysis | μA | 2.426 |

## 5. Conclusion

In this paper a programmable input mode instrumentation amplifier is presented. Such programmability has been not reported in the literature till now, but may be very desirable when a circuit is to be employed in a universal integrated circuit such as an FPGA or a microcontroller. The instrumentation amplifier is built of two multiple output second order current conveyors and 16 transmission gates. The whole circuit has been designed using AMS 350nm CMOS technology in the Cadence Virtuoso environment. The detailed results of simulation of both multiple output current conveyor and instrumentation amplifier are presented. Input mode selection of instrumentation amplifier is performed through appropriate voltage feed to the control nodes. The power supply of the circuit is 3.3 V and it consumes 9.1 mW of power. The IA exhibits a ± 1.68 V linear input range in the voltage mode and ± 250 μA in the current mode. A 3 dB passband of the circuit is located above 11 MHz. The amplifier works in class A, so its current supply is almost constant and does not cause noise disturbing precision analogue circuits working nearby. A dynamic range of the amplifier is equal to 77.3 dB in the input voltage mode and 85.1 dB in the current one. Extensive MC simulations have also been performed. All the results confirm usability of the proposed instrumentation amplifier in real applications in medium precision range applications with the resolution in a range of 10 bits.

## References

[1] Pandey, N., Nand, D., Pandey, R. (2016). Generalised operational floating current conveyor based instrumentation amplifier. *IET Circuits Devices Syst.*, 10(3), 209–219.

[2] Cini, U., Arslan, E. (2015). A High Gain and Low-Offset Current-Mode Instrumentation Amplifier Using Differential Difference Current Conveyors. *Proc of IEEE Int. Conf. on El. Cir. and Syst.*, Egypt, 69–72.

[3] Schaffer, V., Snoeij, M., Ivanov, M. Trifonov D. (2009). A 36 V Programmable Instrumentation Amplifier With Sub-20 V Offset and a CMRR in Excess of 120 dB at All Gain Settings. *IEEE Journal of Solid-State Circuits*, 44(7), 2036–2046.

[4] Tang, A. (2005). Enhanced Programmable Instrumentation Amplifier. *Proc of IEEE Sensors Conf.*, Irvine, USA, 955–958, CD-ROM.

[5] Vyroubal, D. (1990). Instrumentation Amplifier with Digital Gain Programming and Common-Mode Rejection Trim. *IEEE Trans. On Instr. and Meas.*, 39(4), 588–593.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 79–89.

[6] Sedra, A., Smith, K. (1970). A second-generation current conveyor and its applications. *IEEE Transactions on Circuit Theory*, 17(1), 132–134.

[7] Ismail, A.M., Soliman, A.M. (2000). Low-power CMOS current conveyor. *Electronics Letters*, 36(1), 7–8.

[8] Horng, J. W., Hou, C.L., Chang, C.M.(2008). Multi-input differential current conveyor, CMOS realisation and application. *IET Circuits, Devices & Systems*, 2(6), 469–475.

[9] Becvar, D., Vrba, K., Zeman, V., Musil, V. (2000). Novel universal active block: a universal current conveyor. *Proc. of IEEE Int. Symp. on Circuits and Systems*. Geneve, Switzerland, 471–474.

[10] Fani, R., Farshidi, E. (2013). New systematic two-graph-based approach of active filters employing multiple output current controlled conveyors. *IET Circuits, Devices & Systems*, 7(6), 326–336.

[11] Pankiewicz, B. (2016). Multiple output CMOS current amplifier. *Bulletin of the Polish Academy of Sciences, Technical Sciences*, 64(2), 301–306.

[12] Pelgrom, M.J.M., Duinmaijer, A.C.J., Welbers, A.P.G. (1989). Matching properties of MOS transistors, *IEEE Journal of Solid-State Circuits*, 24(5), 1433–1439.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 91–100.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# ANALYSIS OF ERRORS OF PIEZOELECTRIC SENSORS USED IN WEAPON STABILIZERS

**Igor Korobiichuk**

*Industrial Research Institute for Automation and Measurements PIAP, Al. Jerozolimskie 202, 02-486 Warsaw, Poland*
*(✉ ikorobiichuk@piap.pl, +48 516 593 540)*

**Abstract**

Effectiveness of operation of a weapon stabilization system is largely dependent on the choice of a sensor, *i.e.* an accelerometer. The paper identifies and examines fundamental errors of piezoelectric accelerometers and offers measures for their reduction. Errors of a weapon stabilizer piezoelectric sensor have been calculated. The instrumental measurement error does not exceed $0.1 \times 10^{-5}$ m/s$^2$. The errors caused by the method of attachment to the base, different noise sources and zero point drift can be mitigated by the design features of piezoelectric sensors used in weapon stabilizers.

Keywords: acceleration, piezoelectric sensor, errors, weapon stabilizer.



## 1. Introduction

Today, there are several types of sensors for *Weapon Stabilization System* (WSS). Each of them has its advantages and disadvantages. Leading technical universities in Ukraine, Poland, USA, Japan, Germany, Russia and other world's leading countries develop new models of WSS accelerometers and increase their accuracy. The paper presents basic types of accelerometers.

As shown by the analysis of WSS accelerometers, the achievable accuracy of aviation accelerometric measurements is currently $(2–10) \times 10^{-5}$ m/s$^2$ [1–3]. However, in order to satisfy tasks performed by vehicle weapon stabilizers, accelerometry requires a substantially improved accuracy and speed of measurements. It stems primarily from the necessity for improving the accuracy of accelerometers, developing methods for automatic compensation of acceleration measurement errors, improving the mathematical model of WSS, and solving the problem of filtration of perturbations in the WSS accelerometer output signal [4, 5].

The accuracy of a modern WSS is mainly limited by the accuracy of its accelerometer [6].

Ultimately, we can conclude that attaining a WSS accelerometer accuracy of $1 \times 10^{-5}$ m/s$^2$ is currently crucial for a substantial improvement of accuracy of modern WSSs.

Well-known WSS sensors are characterized by the above mentioned advantages but also by their significant disadvantages (Table. 1), which mainly include:

1) a low measurement accuracy ($(2–10) \times 10^{-5}$ m/s$^2$);
2) the necessity for applying a filtering procedure to the WSS sensor output signal;
3) instability of the static gear ratio of WSS accelerometers caused by changes in the properties of their structural elements;
4) a low speed and inability of in-line data processing, and others.

These disadvantages may be overcome, if a piezoelectric sensor is used as a WSS accelerometer [6]. The research in this field is reasonable, since piezoelectric sensors are

Article history: received on Apr. 24, 2016; accepted on Aug. 28, 2016; available online on Mar. 15, 2017; DOI: 10.1515/mms-2017-0001.

currently the best sensors to use in inertial navigation systems and ballistic missile control systems. These devices have been constructed for use in complex dynamic conditions (an axis acceleration of over 50 g; a temperature range: – 80 ... + 200°C; an air pressure in an unpressurized box: 700...800 mm Hg at the surface of the Earth and $10^{-6}$ mm Hg at an altitude of 200 km). Dynamic conditions are less strict when using a *piezoelectric sensor* (PS) as an element of WSS in vehicles [6]. That is why the authors decided to perform more thorough research focusing on the advisability of using piezoelectric accelerometers in WSSs.

Table 1. Comparative analysis of existing WSS accelerometers.

| Type | Denotation | Function | Acceleration measurement accuracy, m/s² | Sensitivity threshold, × 10–5 m/s² | Disadvantages |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| Quartz | GAL | Compensation of acceleration torque by torsion of elastic thread that carries horizontal pendulum | $8 \times 10^{-5}$ | 0.3 | Big time constant; Insufficient speed; Low accuracy; Low sensitivity |
| | GI 1/1 | | $6 \times 10^{-5}$ | 0.1 | |
| | Chekan-AM | | $6 \times 10^{-5}$ | 0.1 | |
| | GRIN-2000/M | | $5 \times 10^{-5}$ | 0.2 | |
| Spring | GSS | Compensation of torque by vertical spring | $10 \times 10^{-5}$ | 0.2 | Hard-to-predict drift of elastic properties of string element; Low accuracy |
| | L-R-S | | | 0.1 | |
| Magnetic | Bell BGM-2, Bell VM-IX, Autonetics VM-7G, MAG-1M, GT-1A, GT-2A | Compensation of acceleration torque by magnetic or electromagnetic spring | $8 \times 10^{-5}$ | 0.2 | Instability of magnetic properties of permanent magnet; Insufficient speed; Low accuracy |
| String | GSD-M | Change of string vibration frequency is directly proportional to change of acceleration | $8 \times 10^{-5}$ | 0.1 | Instability of elastic properties of string; Possibility of resonance; Insufficient speed and accuracy |
| | GRAVITON-M | | $5 \times 10^{-5}$ | 0.1 | |
| Gyroscopic | PIGA 16, 25 | Turn of platform at an angle sufficient to create a gyroscopic torque that balances pendulosity along input axis of a device | $3 \times 10^{-5}$ | 0.1 | High net costs; Structural complexity; Insufficient speed and accuracy |
| | Gyro accelerometer | | $2 \times 10^{-5}$ | 0.1 | |

The literature sources [7–12] neither analyze the basic errors, nor study the main characteristics of new PSs [6]. Therefore, the aim of this section is to provide an analysis and necessary study.

The task is to analyze the methodological WSS errors, determine the composition and structure of PS errors, identify the major errors of a new PS and suggest ways to reduce them.

Available vehicle WSSs, which apply quartz, magnetic, spring, string and gyroscopic accelerometers can provide an acceleration measurement accuracy within $(2–10) \times 10^{-5}$ m/s$^2$. However, an accuracy of WSS accelerometers is required to be $1 \times 10^{-5}$ m/s$^2$ for effective practical application of WSS [6].

The summarized shortcomings of the existing accelerometers for WSSs are eliminated completely or partially due to the use of a piezoelectric element as the WSS accelerometer. WSS contains a piezoelectric element, sensors for determining the object's speed and location coordinates, and current height sensors. Their outputs are connected to a mobile microcontroller

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 91–100.

of objects. The piezoelectric accelerometer is located on a double-axis platform, stabilizing its sensitivity axis vertically. The sensor of the WSS piezoelectric sensing element consists of a piezoelectric element, working on the compression-stretching strain, with insulators on its both ends and inertial mass. The piezoelectric element is a multi-layer structure (piezo-pack) consisting of crystal lithium niobate layers with antiparallel polarization and electrodes separated by connecting layers.

## 2. Piezoelectric sensor errors

To analyze the PS errors, the following classification should be introduced: by the error factors (the methodical factor, caused by an imperfect measurement method or a mismatched model, and the instrumental factor, caused by measuring device properties), by their effects (static and dynamic); by their repeatability (the random errors, varying randomly in sign and value during repeated measurements of the same value, and the systematic errors, remaining during the same measurements either constant or varying according to expectations) [13, 3].

### *2.1. Instrumental errors*

The instrumental PS error is determined as the sum of errors of all values that directly affect the final output of accelerometer [13].

The basic working formula of converting acceleration to voltage is as follows:

$$U_{out} = \frac{d_{ij} \cdot m \cdot g_z}{C_{PS}}, \tag{1}$$

where: $U_{out}$ is the output *PS* voltage; $g_z$ is the *gravitational acceleration* (GA); $d_{ij}$ is the piezoelectric modulus; $m$ is the mass of *PS* and *IM*; $C_{PS}$ is the electric capacity of *PS*.

The true value of GA is determined by the formula:

$$g_z = \frac{U_{out} \cdot C_{PS}}{d_{ij} \cdot m}. \tag{2}$$

The relative error of output signal equals to the sum of multiplications of relative parameter errors by parameter exponents:

$$\frac{\Delta g_z}{g_z} = \frac{\Delta U_{out}}{U_{out}} + \frac{\Delta C_{PS}}{C_{PS}} - \frac{\Delta d_{ij}}{d_{ij}} - \frac{\Delta m}{m}. \tag{3}$$

Let us consider each component of an error:

1) To calculate the piezoelectric modulus variation error, it is worth mentioning that the new PS is made of lithium niobate. At a temperature variation the piezoelectric modulus changes according to the law:

$$\Delta d_{ij} = d_{ij} \alpha_{tc} \cdot \Delta t, \tag{4}$$

where $\alpha_{tc}$ is the temperature coefficient of linear expansion of quartz, $\Delta t$ is the temperature variation value.

The temperature variation error of the piezoelectric modulus:

$$\left( \frac{\Delta d_{ij}}{d_{ij}} \right) = \alpha_{tc} \Delta t. \tag{5}$$

For lithium niobate $\alpha = 0.59 \cdot 10^{-60} \mathrm{C}^{-1}$ [14, 15], then:

$$\left( \frac{\Delta d_{ij}}{d_{ij}} \right) = 0.59 \cdot 10^{-6} \cdot 1 = 0.59 \cdot 10^{-6}. \tag{6}$$

2) To calculate the electrical capacitance variation error, we should use the formula:

$$C_{PS} = \frac{\varepsilon \cdot S}{d}, \tag{7}$$

where: $\varepsilon$ is the lithium niobate dielectric constant; $S$ is the PS area; $d$ is the PS height.
As can be seen from (7), the electrical capacitance error depends on the dielectric constant variation and the area affected by the gravitational acceleration. In accordance with the characteristics of dependence of lithium niobate dielectric constant variation on temperature variation, $\varepsilon$ varies by 0.5% for the temperature change from 0ºC to +500ºC. For 1ºC it is 0.001%. Consequently, the dielectric constant variation error is:

$$\left(\frac{\Delta\varepsilon}{\varepsilon}\right) == 1 \cdot 10^{-5}. \tag{8}$$

The relative PS area variation error $\frac{\Delta S}{S}$:

$$\frac{\Delta S}{S} = \frac{\Delta b}{b} + \frac{\Delta l}{l}, \tag{9}$$

where: $b = 20 \cdot 10^{-3}$ m and $l = 25 \cdot 10^{-3}$ m are the width and length of PS; respectively; $\Delta l$, $\Delta b = 0.8$ mkm are the tolerances for PS area sides.
Then:

$$\frac{\Delta S}{S} = \frac{0.8 \cdot 10^{-6}}{20 \cdot 10^{-3}} + \frac{0.8 \cdot 10^{-6}}{25 \cdot 10^{-3}} = 0.72 \cdot 10^{-4}. \tag{10}$$

The PS height variation error $\frac{\Delta d}{d}$, when $\Delta d = 0.3$mkm, is:

$$\frac{\Delta d}{d} = \frac{0.3 \cdot 10^{-6}}{5 \cdot 10^{-3}} = 0.6 \cdot 10^{-4}. \tag{11}$$

Thus, the electrical capacitance variation error is equal to:

$$\frac{\Delta C_{PS}}{C_{PS}} = \frac{\Delta\varepsilon}{\varepsilon} + \frac{\Delta S}{S} - \frac{\Delta d}{d} = 0.1 \cdot 10^{-4} + 0.72 \cdot 10^{-4} - 0.6 \cdot 10^{-4} = 0.22 \cdot 10^{-4}. \tag{12}$$

3) To calculate the sensor mass variation error, we should use the formula:

$$m = \rho \cdot V, \tag{13}$$

where: $\rho$ is the lithium niobate density; $V$ is the PS volume; $d$ is the PS height.
The PS density variation error mainly depends on the ambient temperature, so, by an analogy to the piezoelectric modulus variation error, we obtain:

$$\left(\frac{\Delta\rho}{\rho}\right) = \alpha_{tc} \Delta t, \tag{14}$$

where $\alpha_{tc} = 0.59 \cdot 10^{-60} \mathrm{C}^{-1}$ [14, 15] is the temperature coefficient of linear expansion of quartz, $\Delta t$ is the temperature variation value.
The relative PS density variation error:

$$\left(\frac{\Delta\rho}{\rho}\right) = 0.59 \cdot 10^{-6} \cdot 1 = 0.59 \cdot 10^{-6}. \tag{15}$$

The PS volume variation error is calculated as follows:

$$\left(\frac{\Delta V}{V}\right) = \frac{\Delta S}{S} + \frac{\Delta d}{d} = 0.72 \cdot 10^{-4} + 0.6 \cdot 10^{-4} = 1.32 \cdot 10^{-4}. \tag{16}$$

Thus, the PS mass variation error is equal to:

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 91–100.

$$\frac{\Delta m}{m} = \frac{\Delta \rho}{\rho} + \frac{\Delta S}{S} + \frac{\Delta d}{d} = 0.59 \cdot 10^{-6} + 0.72 \cdot 10^{-4} + 0.6 \cdot 10^{-4} = 1.32 \cdot 10^{-4}. \quad (17)$$

4) The voltage variation error is determined according to the following consideration. Since the maximum PS instrumental error does not exceed 0.1 mGal (much lesser than the PS cumulative error of 1 mGal), *i.e* $1 \cdot 10^{-5}$ m/s$^2$, we obtain:

$$\frac{\Delta U_{out}}{U_{out}} = \frac{\Delta g_z}{g_z} - \frac{\Delta C_{PS}}{C_{PS}} + \frac{\Delta d_{ij}}{d_{ij}} + \frac{\Delta m}{m} = 0.01 \cdot 10^{-4} - 0.22 \cdot 10^{-4} + 0.0059 \cdot 10^{-4} + \\ + 1.3 \cdot 10^{-4} = 1.1 \cdot 10^{-4}. \quad (18)$$

The error values are summarized in Table 2.

Table 2. The PS instrumental errors.

| № | Components of instrumental error value | Error value |
|---|---|---|
| 1 | Voltage variation, $\frac{\Delta U}{U}$ | $1.1 \cdot 10^{-4}$ |
| 2 | Piezoelectric variation, $\frac{\Delta d_{ij}}{d_{ij}}$ | $0.0059 \cdot 10^{-4}$ |
| 3 | PS electrical capacitance variation, $\frac{\Delta C_{PS}}{C_{PS}}$ | $0.22 \cdot 10^{-4}$ |
| 4 | Mass variation, $\frac{\Delta m}{m}$ | $1.3 \cdot 10^{-4}$ |
| Cumulative instrumental error | | $1 \cdot 10^{-6}$ |

### *2.2. Error due to mechanical mounting of piezoelectric accelerometer to base*

Special attention shall be given to the way of mounting PS to the *horizontally stabilized platform* (HSP) or another base. It is typically an elastic coupling (Fig. 1).

Deficiencies in mounting PS to the base (wrong way of mounting) can lead to significant PS errors.

The errors mainly affect the PS frequency response (appearance of resonances). Errors of this type are insignificant at frequencies up to 200 Hz, Otherwise, they significantly affect PS values. There is a diagram of dependence of mounting methods on the base oscillation frequency values (Fig. 2).

There is a general requirement valid for each mounting method: an almost ideal condition of the base surface.

Screwed pins (3 pieces) should be chosen from the diagram in Fig. 2 as the mounting method. This mounting method corresponds to quite a large working range of PS. That is, the error occurs only at measurement frequencies greater than 10 kHz.

Small parts at the polished base surface should be avoided.

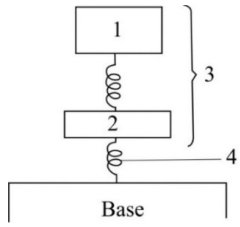



Fig. 1. A mechanical model of PS: 1 – SE; 2 – PS base; 3 – PS; 4 – the mounting method.

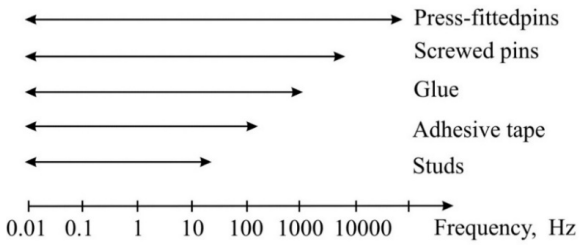



Fig. 2. The dependence of used mounting methods on the operating range of base oscillation frequencies.

### *2.3. Error caused by different noise sources*

There are different types of noise in measurement systems. They are caused by various factors. Therefore, to ensure accurate PS indications, each possible noise should be taken into account and PS should be designed so as to reduce noise to a level at which either it can be neglected in the first approximation, or its impact eliminated.

One of the main types of noise to be primarily reduced is the noise caused by capacitive coupling in PS structure.

The most common way to reduce or eliminate the impact of such noise is to connect the sensor with the measurement circuit by a shielded or coaxial cable. However, this method of avoiding noise involves a relatively small length of shielded cables. Also, the coaxial cable is exposed to moisture which eventually results in decreasing its performance.

The most effective way to deal with the capacitive noise is to use a twisted pair wire, known as the equilibrium (twisted) pair. Since mutual interference in each point of twisted pair results in contraflow, the effect from its action is almost zero at the amplifier input.

After selecting a type of cable, it is necessary to consider the impact of noise on PS figures caused by connecting this cable with the PS construction. Indeed, distortion or displacement of insulation relative to conductors generates a charge motion, mainly under the piezoelectric effects and due to a change of spatial capacity allocation. Such noise can be reduced, if the cable is tightly fixed to the vibrating construction of the perturbed section.

One of circuit designs, enabling a reduction of interference caused by noise, is the use of a protective ring. The high-resistance amplifier input is connected to a low-resistance protection, which is equipotential with respect to the input. Such an amplifier is non-inverting (serves as a buffer). Therefore, its output signal is equal to the input signal, and its output resistance is much less than the input one. The protective ring is directly connected with the amplifier output and forms a low-resistance input to signals from any parasitic coupling.

The acoustic noise also have an impact on PS figures. In particular, this impact is significant during measuring GA. Such noise directly affects PS and place of its mounting to the construction. The level of errors can be illustrated by the fact that the parasitic signal of PS may be about 0.001 g at a sound pressure level of 100 dB.

However, a new PS and a body-base system are well isolated from each other, and it provides a significant resistance to the acoustic noise.

### *2.4. Error caused by piezo-element zero-point shift*

One of disadvantages of accelerometers, which is almost impossible to eliminate, is the zero-point shift or drift. The zero-point shift is reflected in the fact that the PS index always slowly shifts at the same place and in constant conditions (temperature and pressure). So, the readings taken today are different from those taken yesterday. The shift depends on several factors: the ambient temperature, the signal mode and others. The nature of this shift is that the strained

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 91–100.

elastic element of accelerometer (spring, twisted thread or, as in our case, piezo-element) does not precisely follow the law of proportional strain. There is a "fatigue" of the elastic element due to tension, so it is gradually changing its deformation at a constant load [17].

The zero-point shift varies in different systems and for different materials from tenths of mGal to several mGal per day.

The error in a new PS caused by the zero-point shift can be equal to zero for a long time. This is because at low and average temperatures it remains stable, and a feedback loop included in the system constantly returns PS to its original position, compensating for the input load.

### *2.5. Error caused by atmospheric pressure change*

The atmospheric pressure can reduce, to some extent, the load on PS, *i.e.* it is affected only by $G'$, rather than by the whole strength of gravity $G = m{\cdot}g_z$ [8]:

$$G' = mg_z\left(1-\frac{\rho_a}{\rho_m}\right), \tag{19}$$

where: $\rho_a$ is the air density; $\rho_m$ is the sensor's material density.

As the material of a new PS is lithium niobate, its density is $\rho_m$ = 4640 kg/m$^3$, and the air density is $\rho_a$ = 1.2 kg/m$^3$ at a normal atmospheric pressure (101 325 Pa) and temperature 200°C [14, 15]. Substituting these data into the formula (19), we obtain:

$$G' = mg_z\left(1-\frac{1.2}{4640}\right) = (1-0.00026)\cdot mg_z\,. \tag{20}$$

As we see from the formula (20), the gravity decreases by 26 mGal at the required accuracy of 1 mGal ($10^{-5}$ m/s$^2$). This error is calculated for PS working conditions at low altitudes. However, an increase in height entails a decrease in the atmospheric pressure averagely by 11 mm Hg per 100 m and in the ambient temperature by 6°C per 1 km, which causes a change in density of both air and lithium niobate. This phenomenon makes the error unpredictable and difficult for software to calculate and compensate.

There are two ways to ensure stability of PS system to changes in the atmospheric pressure, the first of which is to apply a barometric compensation. This method involves placing PS in a special vacuum chamber that maintains a constant atmospheric pressure. However, it does not protect PS from the adiabatic temperature effect, causing significant errors at a high altitude, and significantly increases its overall size.

Another way to eliminate the impact of atmospheric pressure changes, which is optimal for the PS design as an element of WSS, is pressurizing the PS. That is, the PS and measuring circuit are placed in a sealed enclosure made of a material resistant to changes in the atmospheric pressure and air. PS pressurizing eliminates an influence of error caused by changes in the atmospheric pressure.

### *2.6. Errors of transient (relative to device) angular velocity*

Errors of the transient (relative to the device) angular velocity $\omega_z$ are determined by the formula [3]:

$$\Delta_E = K_{PG}\omega_E\,, \tag{21}$$

$$\delta_E = \frac{\Delta_E}{\alpha_{us}}\cdot 100\%\,, \tag{22}$$

where: $K_{PG}$ is the PS transfer coefficient; $\omega_E$ is the Earth rotational rate; $\alpha_{us}$ is the PS useful signal.

We find the analytical error expression $\Delta_E$, considering that the vertical component of transient angular velocity of the main axis *xOyz* is caused by the Earth's rotation and motion of a vehicle:

$$\omega_z = \omega_E \sin\phi + \frac{v_s}{r} tg\phi , \tag{23}$$

$$v_s = r\dot{\lambda}\cos\phi , \tag{24}$$

$$\frac{v_s}{r} tg\phi = \dot{\lambda}\sin\phi , \tag{25}$$

where $v_s$ is the easterly component of vehicle ground speed; $r$ is the geocentric radius of the Earth; $\dot{\lambda}$ is the longitudinal change rate.

Given (25), the expression (23) can be presented as:

$$\omega_z = (\omega_E + \dot{\lambda})\sin\phi . \tag{26}$$

In general, a vehicle is rotated around the axis *Oz* at the angular velocity $\dot{k}$ and we have:

$$\omega_z = (\omega_E + \dot{\lambda})\sin\phi + \dot{k} , \tag{27}$$

where $k$ is the horizontal course angle, measured clockwise from the north to the longitudinal axis of the object.

Given (27), the expression (21) can be written as:

$$\Delta_E = K_{PG}[(\omega_E + \dot{\lambda})\sin\phi + \dot{k}] . \tag{28}$$

The corresponding average value of the absolute error $\overline{\Delta}_E$ is:

$$(t_2 - t_1)\overline{\Delta}_E = K_{PG}[k(t_2) - k(t_1)] + K_{PG}\int_{t_1}^{t_2} \omega_E \sin\phi(t)dt + K_{PG}\int_{t_1}^{t_2} \dot{\lambda}(t)\sin\phi(t)dt , \tag{29}$$

where $(t_2 - t_1)$ is the averaging interval.

The maximum value of the term $K_{PG}\omega_E \sin\varphi$, corresponding to $\varphi$ = 90°, and the Earth rotational rate $\omega_E = 7.29\cdot10^{-5}$ s$^{-1}$, is $2.92\cdot10^{-5}$ rad [18].

Apparently, the calculation error of the term at stable $\omega_E$ and specified $k$ depends on the calculation error of $\varphi$. Assuming that the calculation error of $K_{PG}\omega_E \sin\varphi$ should not exceed 0.01% = $2.92\cdot10^{-7}$ rad, we can easily calculate that the latitude calculation error should not exceed $0.5^0$.

Note: the latitude calculation error is less than $0.5^0$, if for the averaging interval $(t_2 - t_1)$ the average value $\overline{\sin\varphi}$ is substituted for $\int_{t_1}^{t_2} \sin\varphi(t)dt$ . Besides, since the flights are performed with a constant velocity, the average value $\overline{\varphi}$ corresponds to the midpoint of $(t_2 - t_1)$, and $\overline{\sin\varphi}$ differs from $\sin\overline{\varphi}$ insignificantly, so:

$$K_{PG}\int_{t_1}^{t_2} \omega_E \sin\phi(t)dt = K_{PG}\omega_E \sin\overline{\phi}\,(t_2 - t_1) . \tag{30}$$

WSS sensitivity to the latitude calculation error is maximal at the mid latitude motion of a vehicle. So let us define the term $\dot{\lambda}(t)\sin\varphi$ at $\varphi = 65^0$ and $v_y$ = 234 m/s, $r = 6.4\cdot10^6$ m:

$$\dot{\lambda}(t)\sin\varphi = 7.3\cdot10^{-5}\, c^{-1} . \tag{31}$$

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 91–100.

Consequently, at the predetermined parameters of motion $\dot{\lambda}(t)\sin\varphi$ is equal to the angular velocity of the Earth.

If $\dot{\lambda}(t)$ integral is taken for short time intervals that can be considered constant, we can use the equation:

$$K_{PG}\int_{t_1}^{t_2}\dot{\lambda}(t)\sin\phi(t)dt = K_{PG}[\lambda(t_2)-\lambda(t_1)]\sin\bar{\phi}\,, \tag{32}$$

where $\varphi$ is adjusted as an averaging interval midpoint.

During the test program, a route should be chosen along a parallel (in this case a latitude is almost constant, so the predetermined $\varphi$ can be used in calculations) or a meridian (in this case, a series expansion can be used for relatively crude approximation of $\sin\bar{\varphi}$). When consolidating the flight data for calculating $\bar{\varphi}$, the interval midpoint $(t_2 - t_1)$ should be used.

We write the expression (24) in its final form:

$$\Delta_E = K_{PG}\left(\frac{k(t_2)-k(t_1)}{t_2-t_1}+\omega_E\sin\bar{\phi}+\frac{\lambda(t_2)-\lambda(t_1)}{t_2-t_1}\sin\bar{\phi}\right). \tag{33}$$

Let us calculate $\bar{\Delta}_E$ and $\bar{\delta}_E$ for the above parameters, when $\dot{k}=0$. In this case $\bar{\Delta}_E = 5.8\cdot10^{-5}$ rad = 584 x$10^{-5}$ m/s$^2$ and $\bar{\delta}_E = 2.92\cdot10^{-2}$%.

Therefore, the PS error caused by $\omega_z$ is large compared to other errors. It should be considered when introducing amendments in the equation of WSS motion.

The expression (33) shows that in order to reduce the error of transient angular velocity around the PS axis, we should reduce the transmission factor of instrumentation channel by changing the PS design.

## 3. Conclusions

The research enables to solve a relevant and complex scientific and technical task that is paper identifies and examines the fundamental errors of piezoelectric accelerometers.

Reduction of each type of error is suggested by certain measures (the instrumental error is 0.1×10−5 m/s$^2$).

The composition and structure of PS errors are defined. The main of them are considered and calculated. The instrumental error does not exceed 0.1×10−5 m/s$^2$. The errors caused by the way of attachment to the base, different noise sources and zero point drift can be eliminated by the design features of PS.

## Acknowledgements

This work was supported by the Ministry of Education and Science of Ukraine (grant № 0115U002089).

## References

[1] Lai, A., James, D.A., Hayes, J.P., Harvey, E.C. (2004). Semi-automatic calibration technique using six inertial frames of reference. *Proc. of SPIE × The International Society for Optical Engineering*, 5274, 531–542.

[2] Lakehal, A., Ghemari, Z. (2016). Suggestion for a new design of the piezoresistive accelerometer. *Ferroelectrics*, 493(1), 93–102.

[3] Korobiichuk, I., Bezvesilna, O., Tkachuk, A., Nowicki, M., Szewczyk, R., Shadura, V. (2015). Aviation gravimetric system. *International Journal of Scientific & Engineering Research*, 6(7), 1122–1127.

[4] Liu, Y., Ji, T., *et al*. (2016). Calibration and compensation for accelerometer based on Kalman filter and a six-position method. *Yadian Yu Shengguang/Piezoelectrics and Acoustooptics*, 38(1), 94–98, 110.

[5] Gao, J.M., Zhang, K.B., *et al*. (2015). Temperature characteristics and error compensation for quartz flexible accelerometer. *International Journal of Automation and Computing*, 12(5), 540–550.

[6] Korobiichuk, I. (2016). Mathematical model of precision sensor for an automatic weapons stabilizer system. *Measurement*, *http://dx.doi.org/10.1016/j.measurement.2016.04.017*.

[7] Korobiichuk, I., Bezvesilna, O., *et al*. (2016). Design of piezoelectric gravimeter for automated aviation gravimetric system. *Journal of Automation, Mobile Robotics & Intelligent Systems (JAMRIS)*, 10(1).

[8] Korobiichuk, I., Bezvesilna, O., *et al.* (2016). Piezoelectric gravimeter of the aviation gravimetric system. Advances in Intelligent Systems and Computing 440. Szewczyk, R., Zieliński, C., Kaliczyńska, M. (eds.), *Challenges in Automation, Robotics and Measurement Techniques. Proc. of AUTOMATION-2016*, Warsaw, Poland, 753–763.

[9] Korobiichuk, I., Bezvesilna, O., *et al.* (2015). Stabilization system of aviation gravimeter. *International Journal of Scientific & Engineering Research*, 6(8), 956–959.

[10] Fan, C., Hu, X., *et al*. (2014). Observability analysis of a MEMS INS/GPS integration system with gyroscope G-sensitivity errors. *Sensors*, 14(9), 16003–16016.

[11] Quinchia, A.G., Falco, G., Falletti, E., Dovis, F., Ferrer, C. (2013). A comparison between different error modeling of MEMS applied to GPS/INS integrated systems. *Sensors*, 13(8), 9549–9588.

[12] Karachun, V., Mel'nick, V., Korobiichuk, I., Nowicki, M., Szewczyk, R., Kobzar, S. (2016). The Additional Error of Inertial Sensor Induced by Hypersonic Flight Condition. *Sensors*, 16(3).

[13] Lobanov, V.S., Tarasenko, N.V., *et al.* (2007). Fiber-optic gyros & quartz accelerometers for motion control. *IEEE Aerospace and Electronic Systems Magazine*. 22(4), 23–29.

[14] Guo, Y., Kakimoto, K.I., Ohsato, H. (2005). (Na0.5K0.5)NbO3-LiTaO3 lead-free piezoelectric ceramics. *Materials Letters*, 59(2–3), 241–244.

[15] Tables of fundamental properties of piezoceramic materials manufactured by Ferropiezoelectric Material Division, devices and tools of Research Institute of Physics SFU [electronic resource]. – Access mode, *http://www.piezotech.ru/PKR.htm*.

[16] Arlou, Y.Y., Tsyanenka, D.A., Sinkevich, E.V. (2015). Wideband computationally-effective worst-case model of twisted pair radiation. *Proc. of the International Conference Days on Diffraction*, 14–19.

[17] Meggiolaro, M.A., Castro, J.T.P.D., Góes, R.C.D.O. (2016). Elastoplastic nominal stress effects in the estimation of the notch-tip behavior in tension. *Theoretical and Applied Fracture Mechanics.*

[18] Korobiichuk, I., Koval, A., Nowicki, M., Szewczyk, R. Investigation of the Effect of Gravity Anomalies on the Precession Motion of Single Gyroscope Gravimeter. *Solid State Phenomena*, 251, 139–145.

www.czasopisma.pan.pl PAN www.journals.pan.pl
POLSKA AKADEMIA NAUK

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 101–112.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# KIDNEY SEGMENTATION IN CT DATA USING HYBRID LEVEL-SET METHOD WITH ELLIPSOIDAL SHAPE CONSTRAINTS

**Andrzej Skalski[1)], Katarzyna Heryan[1)], Jacek Jakubowski[2)], Tomasz Drewniak[3)]**

*1) AGH University of Science and Technology, Department of Measurement and Electronics, Al. Mickiewicza 30, Cracow, Poland (✉ skalski@agh.edu.pl, +48 12 617 2828, heryan@agh.edu.pl)*
*2) Rydygier Memorial Hospital, Department of Urology, Os. Zlotej Jesieni 1, 31-826 Cracow, Poland (jacekjakubowski83@gmail.com)*
*3) Specialized Municipal Hospital G. Narutowicz, Department of Urology, Prądnicka 35-37, 31-202 Cracow, Poland (tomdrew@vp.pl)*

**Abstract**

With development of medical diagnostic and imaging techniques the sparing surgeries are facilitated. Renal cancer is one of examples. In order to minimize the amount of healthy kidney removed during the treatment procedure, it is essential to design a system that provides three-dimensional visualization prior to the surgery. The information about location of crucial structures (*e.g.* kidney, renal ureter and arteries) and their mutual spatial arrangement should be delivered to the operator. The introduction of such a system meets both the requirements and expectations of oncological surgeons. In this paper, we present one of the most important steps towards building such a system: a new approach to kidney segmentation from Computed Tomography data. The segmentation is based on the Active Contour Method using the *Level Set* (LS) framework. During the segmentation process the energy functional describing an image is the subject to minimize. The functional proposed in this paper consists of four terms. In contrast to the original approach containing solely the region and boundary terms, the ellipsoidal shape constraint was also introduced. This additional limitation imposed on evolution of the function prevents from leakage to undesired regions. The proposed methodology was tested on 10 Computed Tomography scans from patients diagnosed with renal cancer. The database contained the results of studies performed in several medical centers and on different devices. The average effectiveness of the proposed solution regarding the Dice Coefficient and average Hausdorff distance was equal to 0.862 and 2.37 mm, respectively. Both the qualitative and quantitative evaluations confirm effectiveness of the proposed solution.

Keywords: Level Set method, kidney, CT data, image segmentation, ellipsoid.



## 1. Introduction

Recently, the increase of kidney cancer incidences has been observed. Due to the fact that this type of cancer is frequently diagnosed at an early stage of its development, it is possible to provide the sparing treatment [1]. Its idea is to extract only the tumor lesion with a necessary margin and to remain intact as much of the healthy kidney as possible. This type of treatment requires a precise preoperative planning. For a physician who undertakes the surgery, the information about location of critical structures (kidney, vascular tree, renal pelvis and ureter), their mutual spatial orientation and possible conflicts in the operating field is essential. Therefore, the preoperative imaging, which is in most cases *Computed Tomography* (CT) with a contrast agent, is performed to visualize the structures of interest. Unfortunately, the information presented in this way often appears to be insufficient. It is associated with the kidney spatial arrangement and its geometry, as well as a poor spatial image resolution.

Delivery of a patient-specific three-dimensional model of kidney and other critical structures is indicated as the next step in development of urological oncology [2, 3]. So far only the visualization based solely on the CT data has been used and no reliable automatic tool has been yet proposed. The process of manual preparation of such visualizations is time-consuming

Article history: received on Jun. 07, 2016; accepted on Aug. 28, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0006.

and often inaccurate due to the shift between different data series and insufficient information derived from CT data for such a reconstruction. The optimal approach to facilitate this process seems to be the algorithmic one. In order to make kidney sparing surgery easier, it is important to identify tools for automatic preparation of such a model for each patient. The kidney segmentation, the scope of this paper, is a step on this way. Among the difficulties encountered during the CT data segmentation, there is a slight difference in HU values assigned to the voxels representing the kidney and surrounding tissue. Although several methods for the 3D kidney segmentation have been proposed, they seem to be insufficient for the presented purpose. In general, algorithms of 3D kidney segmentation from the CT data can be divided into 3 main groups:

1. Thresholding, Region Growing and feature extraction methods.
   Lin *et al.* [4] proposed a multistage system. In the first step, location of the spine is used to extract the region candidate. Then, identification the seed point is followed by adaptive region growing. Nedevschi and Mile [5] used a set of texture features derived by Gabor filters and the EM-based image segmentation. Nevertheless, the presented results are coming from 2D slices. These methods are vulnerable to spatial data distribution and dispersion of values; especially present in the case of contrast-enhanced CT data. The data used in our research were of two types: with and without injection of a contrast agent prior to examination.
2. *Atlas and Active Shape/Appearance model* (ASM/AAM) segmentation.
   In [6], the authors proposed a multi-atlas segmentation. The atlas was created using the classical 2D snake segmentation followed by correction made by the radiologist. Xu *et al.* [7] also used the multi-atlas approach. In comparison with the classical multi-atlas strategy, the segmentation was further improved by SIMPLE algorithm with context learning. Chao *et al*. [8] proposed a method for segmentation of kidney compartments. After localization based on Generalized Hough Transform the 3D AAM segmentation is performed. Spiegel *et al.* [9] exploited ASM framework to the kidney segmentation. A crucial correspondence problem in the ASM training phase was solved using the image registration. However, the presented methods require a huge CT database.
3. Active Contours and Level Set segmentation.
   Zolllner *et al.* [10] proposed a 2D Active Contours approach with k-means clustering to provide the initial curve. In [11] the authors introduced the 3D generalization known as Deformable Models. The model is defined by NURBS surfaces constrained by prior statistical information concerning the curvature distribution on the surface. However, this method has a limited robustness to kidney pathologies, which is the case in kidney cancer CT data.
   Khalifa *et al*. used a Level Set method with prior shapes. This methodology was extended by Markov-Gibbs Random Field to model the kidney and surrounding tissues [12]. Huang *et al.* [13] proposed a modification of the Chan-Vese model [14] by using a shape model with kidney variation. Similarly to the methods presented in Section 2, this kind of solution also requires a big database with doctors' manual outlines.

The method proposed in this paper is most congruent with the algorithm presented in [15] for the cardiac MR image segmentation.

## 2. Methodology

In this section, the background of proposed methodology is given. Subsection 2.1 presents a general framework of kidney segmentation based on the Hybrid Level Set method adopted from [16]. The *Hybrid Level Set* (HLS) extended with the shape constraints is described in Subsections 2.2 and 2.3. Since kidneys have an ellipsoidal shape, this feature was added as

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 101–112.

an additional restriction to the evolution equation (Subsection 2.4). In Subsection 2.5, a method of parameter selection is proposed. Finally, in Subsection 2.6 the energy functional minimization process is presented.

### *2.1. Hybrid Level Set*

It was already mentioned in the introduction that our goal is to develop an effective kidney segmentation method from the CT data. The proposed methodology, similarly to that proposed in [15], is based on four major assumptions. First of all, the algorithm should take into account the distribution of values within the object (the region term). The information about the values assigned to the object should be also extended with the boundary information. Moreover, the additional prior knowledge should be introduced, *e.g.* the geometrical model of segmented structure. Finally, smoothness of the resulting kidney contour should be ensured. All of the foregoing assumptions may be included in the *Active Contours* framework.

The idea of image segmentation using *Active Contours* (AC) was introduced in 1988 [17]. Since that time, the researchers have proposed a wide range of AC model modifications. Nonetheless, the AC models are usually based on minimization of the energy functional. The functional is such defined that its smallest value is near or on the target object boundary. In order to solve the minimization problem, a *partial differential equation* (PDE) corresponding to the functional is formulated. From a numerical point of view, the methods for solving this kind of PDE can be divided into three main categories [16]:

- Particle models wherein AC is built of a set of points [17].
- Analytic models in a parametrical form (*e.g.* B-spline [18]).
- *Level Set* (LS) models wherein AC is a zero-level set of functions with an implicit representation defined in a space being by one dimension greater than that of the image [19].

The proposed solution is based on the Level Set model with the boundary information associated with the region information [16].

The preliminary assumptions concerning the proposed solution can be presented in the form of an energy functional $E(C)$ [15]:

$$E(C) = \lambda_1 E_1(C) + \lambda_2 E_2(C) + \lambda_3 E_3(C) + reg\,, \tag{1}$$

where: $E_1(C)$ is the region term; $E_2(C)$ provides information about the boundaries; $E_3(C)$ is the geometrical model of an object and *reg* ensures smoothness of the contour/surface $C$. The parameter $\lambda_i$ is the weight of the $i$-th component impacting the entire functional.

In the LS framework the surface $C$ in the 3-dimensional case is represented as a zero-level set of the function $\Phi(\mathbf{x})$ ($\mathbf{x} = [x_1, x_2, x_3]$) [14, 19]:

$$\begin{cases} C = \{\mathbf{x} \in \Omega : \Phi(\mathbf{x}) = 0\} \\ \Omega_{\text{in}} = \{\mathbf{x} \in \Omega : \Phi(\mathbf{x}) > 0\} \\ \Omega_{\text{out}} = \{\mathbf{x} \in \Omega : \Phi(\mathbf{x}) < 0\} \end{cases}. \tag{2}$$

Using the notation (2), the functional (1) is represented as:

$$E(C) = \lambda_1 E_1(\Phi) + \lambda_2 E_2(\Phi) + \lambda_3 E_3(\Phi) + reg\,. \tag{3}$$

### *2.2. Region-based term*

Contrary to the original approach proposed in [14], here the $E_1(C)$ term is based on the functional component of HLS [16]:

$$E_1(\Phi) = -\int_{\Omega} \big(\mathbf{I}(\mathbf{x}) - \mu\big) H(\Phi) d\Omega \,, \tag{4}$$

where: **I** represents the image data used for segmentation; $H(\Phi)$ is the Heaviside function ($H(\Phi) = 0$ for $\Phi<0$ and $H(\Phi) = 1$ for $\Phi \geq 0$), $\mu$ is a parameter indicating the lower threshold value associated with the object, assuming that the object values are relatively higher than those of the background (the method of determining this parameter is described in Subsection 2.5). In consequence of this way of determining (4), the voxels with values greater than $\mu$ are preferred [16]. In comparison with the original solution proposed in [14], the number of parameters that require determination is reduced from 2 to 1. Also, this approach eliminates the problem of background/object representation. Moreover, in the kidney surroundings there are voxels with very low values (*e.g.* air in a small intestine or colon), similar values (surrounding soft tissues), *e.g.* the liver, and those with very large ones (*e.g.* bones). Due to such a determination of the boundary term (4), the problem associated with dispersion of the background voxel values is eliminated, which is crucial in this case. In our research, $\mu$ parameter was determined prior to segmentation and remained constant during the evolution process.

### *2.3. Edge-based term*

Only if the surrounding structures have similar values and there is a clear boundary between them, the introduction of the image edge information makes the solution more effective. The functional component representing the edge-based term is defined as follows [16]:

$$E_2(\Phi) = \int_{\Omega} \mathbf{g}(\mathbf{x}) \big| \nabla H(\Phi) \big| d\Omega \,, \tag{5}$$

where **g** is the boundary function defined as:

$$\mathbf{g}(\mathbf{x}) = \frac{1}{1 - c\big|\nabla \mathbf{I}(\mathbf{x})\big|^2} \,, \tag{6}$$

where $c$ is a slope control parameter.

### *2.4. Prior kidney shape term*

The major obstacle is that the kidney voxel values are similar to these of the surrounding structures. Moreover, for contrast-enhanced data, sometimes the contrast agent is not propagating as desired, leading to significant dispersion of values within the kidney itself. A similar phenomenon may occur when the image acquisition time is set improperly in relation to the contrast injection.

The proposed solution is based on the 3D generalization of the 2D case presented in [15]. It is modified by adding a functional component representing the object shape. The shape constraint can be provided in several ways. One approach is to use an average shape of the segmented structure with its acceptable variability. Its significant limitation concerns the requirement for a greatly expanded image database describing the kidney and both its geometric and volumetric variations. Another solution that we propose, is to use the feature of kidney similarity to the ellipsoidal shape. An ellipsoid in the three-dimensional space can be described using the quadratic form:

$$a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + a_{12}x_1x_2 + a_{23}x_2x_3 + a_{13}x_1x_3 + a_{14}x_1 + a_{24}x_2 + a_{34}x_3 + a_{44} = 0 \,. \tag{7}$$

Or, in a more compact form:

$$\boldsymbol{a}^T \boldsymbol{z} = 0 \,, \tag{8}$$

where: $\mathbf{a}^T = [a_{11}, a_{22}, a_{33}, a_{12}, a_{23}, a_{13}, a_{14}, a_{24}, a_{34}, a_{44}]$, $\mathbf{z}^T = [x_1^2, x_2^2, x_3^2, x_1x_2, x_2x_3, x_1x_3, x_1, x_2, x_3, 1]$.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 101–112.

In this case, the crucial issue is to ensure that (8) will be indeed an ellipsoid. The conditions for it are as follows [20, 21]:

$$4\left(a_{11}a_{22}+a_{22}a_{33}+a_{11}a_{33}-\frac{1}{4}a_{12}^2-\frac{1}{4}a_{23}^2-\frac{1}{4}a_{13}^2\right)-\left(a_{11}+a_{22}+a_{33}\right)^2>0,\ 4M-L>0. \quad (9)$$

If $4M-L>0$, the vector of searching parameters $\mathbf{a}$ represents an ellipsoid. Based on this, it is possible to formulate an optimization problem: $\min_{\mathbf{a}} \mathbf{a}^T\mathbf{R}\mathbf{a}$ and $\mathbf{a}^T\mathbf{Q}\mathbf{a}=1$, where the matrix $\mathbf{R}$ represents points belonging to the zero-level set of $\Phi$ and $\mathbf{Q}$ is a constraint matrix presented in [20]:

$$\mathbf{Q}=\operatorname{diag}\left(-\frac{1}{2}k\left(\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}+\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\right),-\frac{1}{4}k\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix},\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\right), \quad (10)$$

where $k\geq 4$. More details and a fitting procedure can be found in [20].

After estimation of the vector $\mathbf{a}$, another level-set function $\mathbf{D}$ representing the distance to the ellipsoid surface can be specified. $\mathbf{D}$ can be calculated in two ways:

$$\mathbf{D}_I=d;\ \mathbf{D}_{II}=\begin{cases}d & \text{outside ellipsoid}\\ 0 & \text{inside ellipsoid}\end{cases}, \quad (11)$$

where $d$ is a distance to the ellipsoid equal to $\mathbf{a}^T\mathbf{z}$. $\mathbf{D}_I$ prefers the external surface as the final result, whereas $\mathbf{D}_{II}$ allows for topological changes inside the ellipsoid. By using $\mathbf{D}_{II}$ the inner structure of kidney can be differentiated. The results of the segmentation process depending on choosing either $\mathbf{D}_I$ or $\mathbf{D}_{II}$ are presented in Fig. 1.

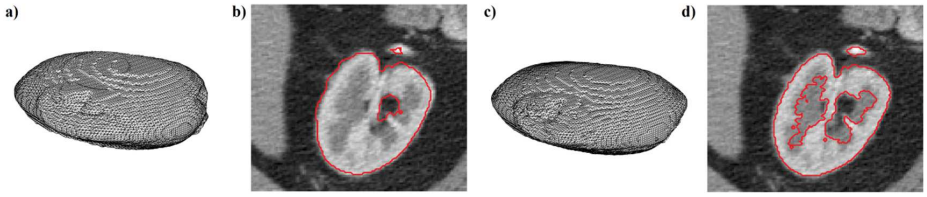



Fig. 1. An example of the results with $\mathbf{D}_I$, (a) and (b), $\mathbf{D}_{II}$ (c), and (d) the formulation used in (12); (a), (c) the final kidney surface; (b), (d) the result (red contour) on the slice through (a) and (c).

Finally, $E_3(C)$ can be determined:

$$E_3(\Phi)=\int_{\Omega}\mathbf{D}^2(\mathbf{x})\left|\nabla H(\Phi)\right|d\Omega. \quad (12)$$

### *2.5. Estimation of μ*

The value of $\mu$ parameter can be obtained by different methods. The simplest possible way is its indication based on thresholding or manual histogram analysis. Automatic results can be achieved by interpretation of normalized histogram function approximation parameters. For this purpose, in this study, the approximation of normalized histogram is performed by the *Gaussian Mixture Model* (GMM) using the iterative Expectation-Maximization algorithm [22].

GMM consists of $k$ Gaussian functions. Each function is parameterized by $\overline{\mathbf{w}}_i$ that represents the position of peak centre and by $\sigma_i$ that controls its width. This can be formulated as follows:

$$GMM = \sum_{i=1}^{k} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left( \frac{-(w-\overline{\mathbf{w}}_i)^2}{2\sigma_i^2} \right), \tag{13}$$

where $w$ represents values assigned to the image voxels. In each iteration over $k$, where the value of $k$ is between 2 (two Gaussian functions used for approximation of the normalized image histogram and differentiation between the background and object) and $K$, an appropriate value of $\overline{\mathbf{w}}_i$ is chosen.

First, among the $\overline{\mathbf{w}}_i$ values the ones whose contributions to approximation exceed 10%, are selected. Secondly, $\overline{\mathbf{w}}_i$ are sorted in the ascending order. If the data are enhanced by a contrast agent, the highest value of $\overline{\mathbf{w}}_i$ is chosen (the value resembling bones and tissues with a contrast). Otherwise, the penultimate value of $\overline{\mathbf{w}}_i$ (the value resembling a soft tissue, in this case also the kidney) is picked out. Next, it is checked whether the value of $\sigma_i^2$ corresponding to the chosen $\overline{\mathbf{w}}_i$ is typical for the component representing the kidney. Too high a value of $\sigma_i^2$ implies iterative continuation of the estimation process with $k = k + 1$ (probably other structures with different values are around the kidney). Finally, $\mu$ is equal to $\overline{\mathbf{w}}_i$ for the chosen $k$. An example of approximation of a normalized image histogram by GMM with $k = 3$ and $k = 4$ is presented in Fig. 2. In some cases fixing the $\mu$ value has to be done manually.

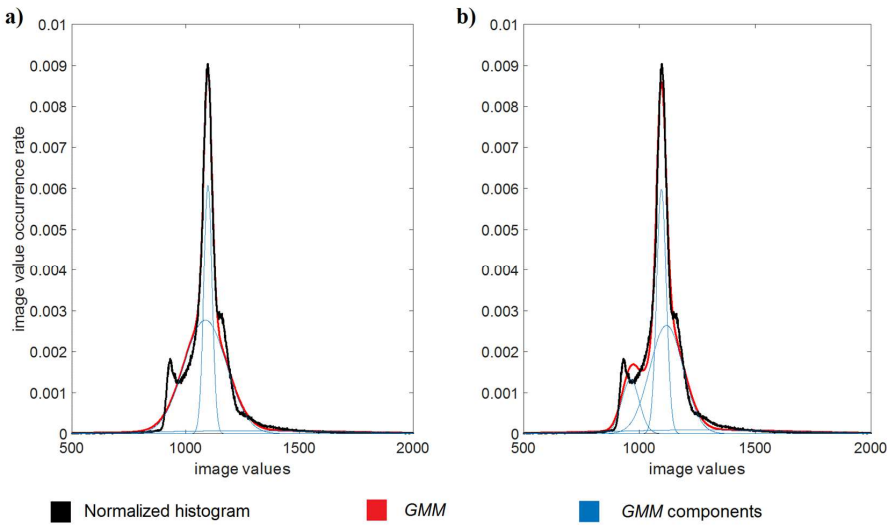



Fig. 2. An example of approximation of a normalized image histogram by GMM with $k = 3$ (a) and $k = 4$ (b). The data are without a contrast agent which implies that the penultimate value of $\overline{\mathbf{w}}_i$ is selected. In the case of (a) two Gaussian functions are meaningful, but the value of $\sigma_i^2$ corresponding to the chosen $\overline{\mathbf{w}}_i$ is too high. In the case of (b) three Gaussian functions are meaningful, and the value of $\sigma_i^2$ is appropriate.

### *2.6. Minimization of energy functional*

The proposed functional (3) can be rewritten as:

$$E(C) = -\lambda_1 \int_{\Omega} \left(\mathbf{I}(\mathbf{x}) - \mu\right) H(\Phi) d\Omega + \lambda_2 \int_{\Omega} \mathbf{g}(\mathbf{x}) \left|\nabla H(\Phi)\right| d\Omega + \lambda_3 \int_{\Omega} \mathbf{D}^2(\mathbf{x}) \left|\nabla H(\Phi)\right| d\Omega . \tag{14}$$

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 101–112.

By minimizing the energy functional with respect to Φ, a PDE determining the evolution of Φ (derived from the Gateaux derivative gradient flow [16]) was defined:

$$\Phi_t = \delta(\Phi)\left[\lambda_1\left(\mathbf{I}(\mathbf{x})-\mu\right)+\lambda_2\mathrm{div}\left(\mathbf{g}(\mathbf{x})\frac{\nabla\Phi}{|\nabla\Phi|}\right)+\lambda_3\mathrm{div}\left(\mathbf{D}^2(\mathbf{x})\frac{\nabla\Phi}{|\nabla\Phi|}\right)\right]. \tag{15}$$

According to [14, 21] $\delta(\Phi)$ can be replaced by $|\nabla\Phi|$:

$$\Phi_t = |\nabla\Phi|\left[\lambda_1\left(\mathbf{I}(\mathbf{x})-\mu\right)+\lambda_2\mathrm{div}\left(\mathbf{g}(\mathbf{x})\frac{\nabla\Phi}{|\nabla\Phi|}\right)+\lambda_3\mathrm{div}\left(\mathbf{D}^2(\mathbf{x})\frac{\nabla\Phi}{|\nabla\Phi|}\right)\right]. \tag{16}$$

Based on the identity $\mathrm{div}\left(\mathbf{g}\vec{\mathbf{f}}\right)=\nabla\mathbf{g}\cdot\vec{\mathbf{f}}+\mathbf{g}\,\mathrm{div}\left(\vec{\mathbf{f}}\right)$, where "$\cdot$" denotes the inner product, (16) can be written as ($\mathbf{g}=\mathbf{g}(\mathbf{x}), \mathbf{I}=\mathbf{I}(\mathbf{x}), \mathbf{D}=\mathbf{D}(\mathbf{x})$):

$$\Phi_t = |\nabla\Phi|\left[\lambda_1\left(\mathbf{I}-\mu\right)+\lambda_2\nabla\mathbf{g}\cdot\frac{\nabla\Phi}{|\nabla\Phi|}+\lambda_2\mathbf{g}\,\mathrm{div}\left(\frac{\nabla\Phi}{|\nabla\Phi|}\right)+\lambda_3\nabla\mathbf{D}^2\cdot\frac{\nabla\Phi}{|\nabla\Phi|}+\lambda_3\mathbf{D}^2\mathrm{div}\left(\frac{\nabla\Phi}{|\nabla\Phi|}\right)\right]. \tag{17}$$

Using notation $\kappa=\mathrm{div}\left(\frac{\nabla\Phi}{|\nabla\Phi|}\right)$, (17) has the following form:

$$\Phi_t = |\nabla\Phi|\left[\lambda_1\left(\mathbf{I}-\mu\right)+\lambda_2\nabla\mathbf{g}\cdot\frac{\nabla\Phi}{|\nabla\Phi|}+\lambda_2\mathbf{g}\kappa+\lambda_3\nabla\mathbf{D}^2\cdot\frac{\nabla\Phi}{|\nabla\Phi|}+\lambda_3\mathbf{D}^2\kappa\right], \tag{18}$$

where $\kappa$ represents the curvature of evolving contour/surface $C$. The objective of this term is ensuring smoothness of the surface. The evolution is performed in the iterative way based on splitting the additive operator. The details of the numerical implementation can be found in [16].

## 3. Experiment and results

The CT dataset used for testing the proposed kidney segmentation framework consisted of 10 patients diagnosed with renal cancer. The data were acquired in different medical centers and therefore the examinations were performed on different devices and using different acquisition protocols. In consequence, in our dataset there are 2 CT scans with and 8 without a contrast agent injected during examination. Furthermore, the spatial resolution of data in the dataset was various (the spacing range: 0.6094–0.8848 mm, the slice thickness range 1–3 mm).

Figure 3 shows the influence of individual components on segmentation of CT data. It can be observed that lack of the ellipsoidal shape constraints on the evolving surface leads to excessive expansion to the neighboring structures. For example (Figs. 3a and 3b), after occurring a topological change, the zero-level set of Φ has propagated to ribs adjacent to the kidney (high values assigned to the voxels). Introduction of an additional boundary term resulted in reduction of divisions within the kidney only to the areas with the largest gradient (Fig. 3e). The strength of an impact of the boundary term can be adjusted via the parameter $\lambda_2$.

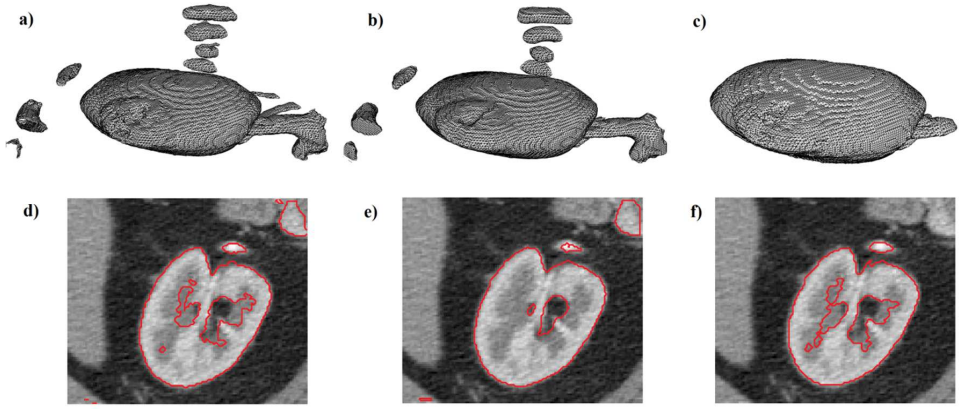



Fig. 3. The influence of energy terms. (a) to (c) visualization of three-dimensional results; (d) to (f) an example of a slice with segmentation results (red contour). (a) and (d) only the area term; (b) and (e) the area and boundary terms; (c) and (f) the area and boundary terms with the ellipsoidal shape constraint for $\mathbf{D}_{II}$ formulation.

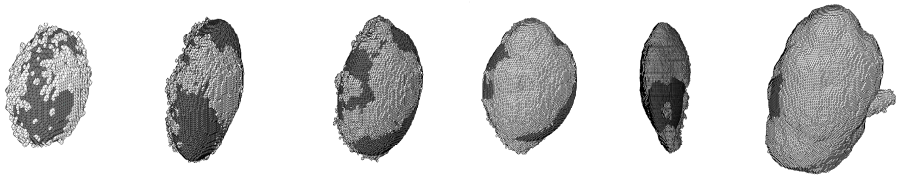


Fig. 4. The zero-level sets of Φ (light grey) and **D** (dark grey) in consecutive iterations. From left to right: iterations: 5, 10, 15, 20, 25, 35.

Additionally, the impact of shape constraint is also presented in Fig. 4. The zero-level sets of estimated shape **D** (dark grey) and function Φ (light grey) are shown on consecutive iterations.

To ensure proper validation of the segmentation results in each case the manual segmentation was performed using ITK-SNAP software [24]. In order to assess effectiveness of the segmentation, the three following measures were used: a typical volumetric measure – the Dice Coefficient (19) [25] and two Hausdorff spatial distance measures [26]: maximum $dH_{\max}$ (20) and average $dH_{\text{avg}}$ (21).

Assuming that $\mathbf{M}_{\text{mc}}$ is a binary mask from manual segmentation and $\mathbf{P}_{\text{mc}}$ is a set of coordinates of its points, $\mathbf{M}_{\text{seg}}$ is a binary mask obtained by the algorithm and $\mathbf{P}_{\text{seg}}$ is a set of coordinates of its points, the aforementioned measures can be formulated as:

$$DICE = \frac{2\left(\mathbf{M}_{\text{mc}} \cap \mathbf{M}_{seg}\right)}{\mathbf{M}_{\text{mc}} \cup \mathbf{M}_{seg}}, \tag{19}$$

$$dH_{\max} = \max\left(h_1\left(\mathbf{P}_{\text{mc}}, \mathbf{P}_{\text{seg}}\right), h_1\left(\mathbf{P}_{\text{seg}}, \mathbf{P}_{\text{mc}}\right)\right), \tag{20}$$

$$dH_{avg} = \max\left(h_2\left(\mathbf{P}_{\text{mc}}, \mathbf{P}_{\text{seg}}\right), h_2\left(\mathbf{P}_{\text{seg}}, \mathbf{P}_{\text{mc}}\right)\right), \tag{21}$$

where: $h_1\left(\mathbf{P}_{\text{mc}}, \mathbf{P}_{\text{seg}}\right) = \max\limits_{p_{mc} \in \mathbf{P}_{mc}} \min\limits_{p_{seg} \in \mathbf{P}_{seg}} \left\| p_{mc} - p_{seg} \right\|$, $h_2\left(\mathbf{P}_{\text{mc}}, \mathbf{P}_{\text{seg}}\right) = \text{mean}\limits_{p_{mc} \in \mathbf{P}_{mc}} \min\limits_{p_{seg} \in \mathbf{P}_{seg}} \left\| p_{mc} - p_{seg} \right\|$, where $\| \ \|$ denotes the Euclidian distance.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 101–112.

To calculate the Hausdorff distances, the Euclidian distances between two subsets $\mathbf{P}_{mc}$ and $\mathbf{P}_{seg}$ are determined. $h_1$ and $h_2$ are formulated as the greatest and the average of all distances, respectively. In general, $dH_{avg}$ shows the difference between manual and automatic segmentation regarding the distance. Since it is averaged over all points, it is known to be more robust and less prone to the noise characteristic of CT data. However, $dH_{max}$ exhibits outliers. This means that even one outlier can contribute to a huge value of $dH_{max}$. The connection of all three presented measures enables to fully assess effectiveness of the segmentation.

The obtained results are presented in Table 1. The lower value of the *DICE* coefficient is caused by the accuracy of manual outlines related to the complex kidney structure. The manual outlines were performed slice by slice what led to further complications in the upper and lower parts of the kidney. This is a consequence of the finite spatial resolution, which hinders specification of the first and the last slices belonging to the kidney (a common phenomenon for spherical and ellipsoidal structures). Furthermore, the manual outlines included only the structure of kidney itself, without arteries, veins and the pelvicaliceal system. A segmentation algorithm sometimes appended these structures to the segmentation results, especially if the introduced contrast agent was located inside them. The obtained mean value of *DICE* coefficient is equal to 0.862 and is on a comparable level to those presented in the literature. The high value of maximum Hausdorff distance results from the fact that among the segmentation results there were vascular structures located within the kidney area (in particular, Figs. 5a, c, g, i).

Table 1. The segmentation effectiveness regarding the Dice and Hausdorff distance measures.

| | ***DICE*** | $dH_{max}$ **[mm]** | $dH_{avg}$ **[mm]** |
|---|---|---|---|
| Mean | 0.862 | 19.63 | 2.37 |
| Standard deviation | 0.039 | 4.99 | 0.62 |

The segmentation results in the form of 3D hull are shown in Fig. 5. In addition, Fig. 6 presents a comparison of the manual outlines and segmentation results on the 2D transversal planes.

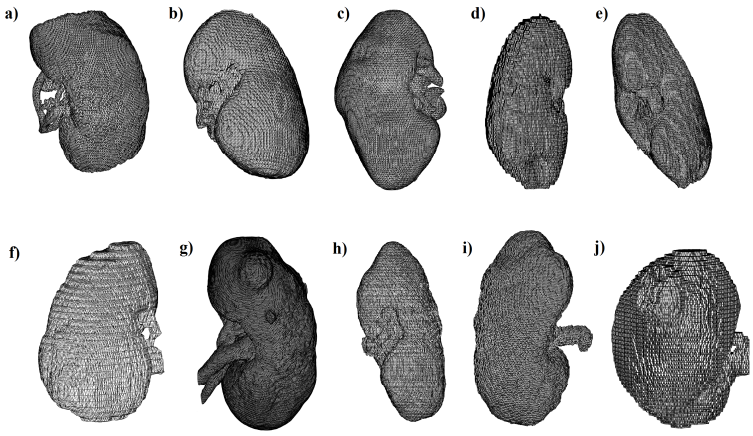



Fig. 5. The segmentation results presented as a three-dimensional model from a different angle of view and for different patients.

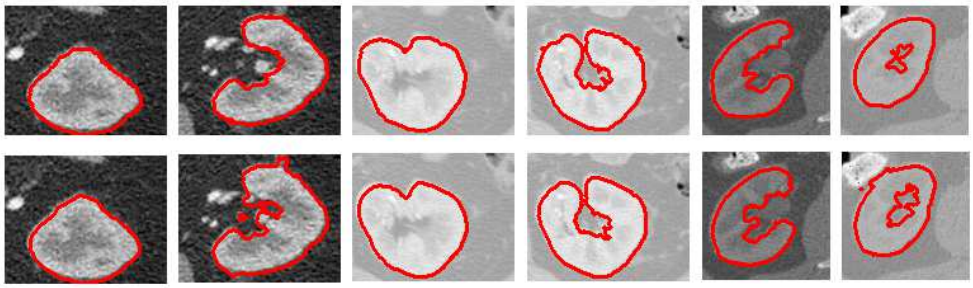


Fig. 6. A comparison of the segmentation results (second row) and manual contours (first row) presented on the transversal planes for 3 patients.

## 4. Conclusion

In this paper, a novel kidney segmentation algorithm using the Level Set framework with ellipsoidal shape constraints was presented. The proposed solution not only takes into account the information about the region and boundary terms, but also the constraints dedicated to the characteristic kidney shape. Thereby, an excessive growth towards undesired structures is restricted. The impact of individual energy functional terms on the segmentation process was also presented. Moreover, based on the Mixture of Gaussians, a simple and effective method for estimation of $\mu$ parameter was described (Subsection 2.5).

The proposed methodology was tested on CT scans carried out on renal cancer patients, both with and without injection of a contrast agent prior to examination. Both the quantitative (Table 1: the *DICE* coefficient equal to $0.862 \pm 0.039$ and the average Hausdorff distance equal to $2.37 \pm 0.62$mm) and qualitative evaluations presented in Fig. 5 and Fig. 6 confirm effectiveness of the proposed solution. It may constitute a significant part of the support system for planning minimally invasive treatments in oncological urology.

Further developing path of the proposed algorithm can cover examination of cross-sectional data in real-time imaging modalities used intraoperatively. The CT/MR-ultrasound fusion image guidance combined with a tracking system can assist percutaneous diagnostic and ablation procedures. Another possibility is augmentation of the laparoscopic picture by guidance provided by CT reconstructions where automatic segmentation is also required.

## Acknowledgements

The work was supported by the Ministry of Science and Higher Education, Poland (Dean Grants, statutory activity).

## References

[1] Klatte, T., *et al*. (2015). A literature review of renal surgical anatomy and surgical strategies for partial nephrectomy. *European urology*, 68(6), 980–992.

[2] Shao, P., *et al*. (2013). Application of a vasculature model and standardization of the renal hilar approach in laparoscopic partial nephrectomy for precise segmental artery clamping. *European Urology*, 63(6), 1072–1081.

[3] Bugajska, K., *et al*. (2015) The renal vessel segmentation for facilitation of partial nephrectomy. *IEEE SPA 2015: Signal Processing: Algorithms, Architectures, Arrangements and Applications*, 50–55.

[4] Lin, D.T., Lei, C.C., Hung, S.W. (2006). Computer-aided kidney segmentation on abdominal CT images. *IEEE Transactions on Information Technology in Biomedicine*, 10(1), 59–65.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 101–112.

[5] Nedevschi, S., Ciurte, A., Mile, G. (2008). Kidney CT image segmentation using multi-feature EM algorithm, based on Gabor filters. *4th International Conference on Intelligent Computer Communication and Processing*, ICCP 2008, 283–286.

[6] Yang, G., *et al*. (2014). Automatic kidney segmentation in CT images based on multi-atlas image registration. *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC), 5538–5541.

[7] Xu, Z., *et al.* (2015). Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning. *Medical image analysis*, 24(1), 18–27.

[8] Chao, J., *et al*. (2016). 3D Fast Automatic Segmentation of Kidney Based on Modified AAM and Random Forest. Accepted to *IEEE Transaction on medical Imaging.*

[9] Spiegel, M., *et al*. (2009). Segmentation of kidneys using a new active shape model generation technique based on non-rigid image registration. *Computerized Medical Imaging and Graphics*, 33(1), 29–39.

[10] Zollner, F.G., Kocinski, M., Rorvik, J. Lundervold. (2007). Towards Automatically Assessment of Kidney Volume from 3D DCE-MRI Time Courses using Active Contours. *Proc. Intl. Soc. Mag. Reson. Med.*, 15.

[11] Tsagaan, B.S.A., Kobatake, H., Miyakawa, K., Hanzawa, Y. (2001). Segmentation of kidney by using a deformable model. *Image Processing, 2001, Proceedings, 2001 International Conference on* 3, 1059–1062.

[12] Khalifa, F., *et al.* (2011). A new deformable model-based segmentation approach for accurate extraction of the kidney from abdominal CT images. *18th IEEE International Conference on Image Processing* (ICIP), 3393–3396.

[13] Huang, Y., *et al.* (2009). Multiphase level set with multi dynamic shape models on kidney segmentation of CT image. *Biomedical Circuits and Systems Conference*, BioCAS, 141–144.

[14] Chan, T.F., Vese, L.A. (2001). Active contours without edges. *IEEE transactions on Image processing*, 10(2), 266–277.

[15] Pluempitiwiriyawej, C., *et al*. (2005). STACS: New active contour scheme for cardiac MR image segmentation. *IEEE Transactions on Medical Imaging*, 24(5), 593–603.

[16] Zhang, Y., *et al*. (2008). Medical image segmentation using new hybrid level-set method. *BioMedical Visualization, 2008, MEDIVIS'08. Fifth International Conference*, 71–76.

[17] Kass, M., Witkin, A., Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, 1(4), 321–331.

[18] Brigger, P., Hoeg, J., Unser, M. (2000). B-spline snakes: a flexible tool for parametric contour detection. *IEEE Transactions on Image Processing*, 9(9), 1484–1496.

[19] Osher, S., Fedkiw, R. (2006). *Level set methods and dynamic implicit surfaces*. Springer Science & Business Media.

[20] Hunyadi, L., Vajk, I. (2014). Constrained quadratic errors-in-variables fitting. *The Visual Computer*, 30(12), 1347–1358.

[21] Li, Q., Griffiths, J.G. (2004). Least squares ellipsoid specific fitting. *Geometric Modeling and Processing*, 335–340.

[22] McLachlan, G., Peel, D. (2000). *Finite Mixture Models. Hoboken*. NJ: John Wiley & Sons, Inc.

[23] Vese, L.A., Chan, T.F. (2002). A multiphase level set framework for image segmentation using the Mumford and Shah model. *International journal of computer vision*, 50(3), 271–293.

[24] Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3), 1116–28.

[25] Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302.

[26] Dubuisson, M.P., *et al*. (1994). A modified Hausdorff distance for object matching. *ICPR94*, 566–68.

www.czasopisma.pan.pl PAN www.journals.pan.pl
POLSKA AKADEMIA NAUK

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 113–125.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# NANOSATELLITE ATTITUDE ESTIMATION FROM VECTOR MEASUREMENTS USING SVD-AIDED UKF ALGORITHM

**Demet Cilden[1)], Halil E. Soken[2)], Chingiz Hajiyev[1)]**

*1) Istanbul Technical University, Faculty of Aeronautics and Astronautics, Maslak, 34469, Istanbul, Turkey*
*(✉ cilden@itu.edu.tr, +90 212 285 3122, cingiz@itu.edu.tr)*
*2) Japan Aerospace Exploration Agency (JAXA), Institute of Space and Astronautical Science (ISAS), Sagamihara, Japan*
*(ersin_soken@ac.jaxa.jp)*

**Abstract**

The integrated *Singular Value Decomposition* (SVD) and *Unscented Kalman Filter* (UKF) method can recursively estimate the attitude and attitude rates of a nanosatellite. At first, Wahba's loss function is minimized using the SVD and the optimal attitude angles are determined on the basis of the magnetometer and Sun sensor measurements. Then, the UKF makes use of the SVD's attitude estimates as measurement results and provides more accurate attitude information as well as the attitude rate estimates. The elements of "Rotation angle error covariance matrix" calculated for the SVD estimations are used in the UKF as the measurement noise covariance values. The algorithm is compared with the SVD and UKF only methods for estimating the attitude from vector measurements. Possible algorithm switching ideas are discussed especially for the eclipse period, when the Sun sensor measurements are not available.

Keywords: attitude estimation, nanosatellite, UKF, SVD, SVD-Aided UKF.



## 1. Introduction

Sun sensors and magnetometers are common attitude sensors for nanosatellite missions; they are cheap, simple, light and available as commercial of-the-shelf equipment [1, 2]. However, the overall achievable attitude determination accuracy is limited with these sensors mainly as a result of their inherent limitations and unavailability of the Sun sensor measurements when the satellite is in the eclipse.

Attitude estimation with magnetometer and Sun sensor measurements has been addressed in many research works and various algorithms that intend to improve the estimation accuracy have been proposed. A basic solution is using a Kalman filtering algorithm for integrating the measurements under the propagation model of the satellite dynamics and estimating the satellite attitude possibly along with the sensor biases. For example, in [2] two filtering algorithms are proposed, both based on the multiplicative *Extended Kalman Filter* (EKF). The first algorithm is used for estimation of attitude quaternions, gyro biases and Sun sensor calibration parameters, whereas the second one estimates only the quaternions and gyro biases excluding the Sun sensor calibration parameters. The main drawback of both algorithms is a degradation in the estimation results when the satellite is in its eclipse so the Sun sensor data are not available. A similar phenomenon can be seen in [1] for the *Unscented Kalman Filter* (UKF) estimations. Another approach to the nanosatellite attitude estimation is to determine the attitude using a single-frame attitude estimator. This method is based on computing Sun and magnetic field vectors in the reference frame and measuring the same vectors in the body coordinate system. Then, a deterministic method such as the TRIAD (*two-vector algorithm*) or an optimization method such as the QUEST can be used for the attitude estimation [3]. A drawback of these methods

Article history: received on Mar. 15, 2016; accepted on Sep. 24, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0011.

is that they are based only on measurements; they do not use any knowledge about the satellite dynamics. Attitude estimation methods, which take the advantage of the system's mathematical model, may significantly increase the attitude estimation accuracy. In [4], the Sun-eclipse phases are considered to use both traditional and non-traditional methods depending on whether the Sun sensor is operational or not. In the Sun sensor operational mode, the Gauss-Newton method enables to obtain the quaternion estimates for using in EKF. In the eclipse mode, only the traditional EKF is used. The measurement covariance values in EKF are not provided by the deterministic method to the filter and they are selected. This leads to some jumps in the filter even outside of the eclipse. If the variance values of the first method are used as the measurement noise covariance ones in EKF, the filter will have to compensate these errors.

The traditional approaches to designing a *Kalman Filter* (KF) for the satellite attitude estimation use nonlinear measurements of reference directions (*e.g.* the Sun direction) [1, 5–7]. The measurement models in the filter are based on nonlinear models of the reference directions so the measurements and states are related by nonlinear equations. In the approach based on linear measurements the attitude angles are found first by using the vector measurements and then a suitable single-frame attitude estimation method [3]. Then, these attitude estimates are used as the measurement results within the KF. The filter measurement model is linear in this case, since the single-frame attitude estimator provides directly the states themselves as measurements. We may name such algorithms as "single-frame estimator-aided attitude filtering".

An earlier study on single-frame estimator-aided attitude filtering was carried out in [8]. In this study the authors integrate the algebraic method (TRIAD) and the EKF algorithms to estimate the attitude angles and angular velocities. The magnetometers, Sun sensors, and horizon scanners/sensors are used as measurement devices and three different two-vector algorithms based on the Earth's magnetic field, Sun, and nadir vectors are proposed. An EKF is designed to obtain the satellite's angular motion parameters with the desired accuracy. The measurement inputs for the EKF are the attitude estimates of the two-vector algorithms. Interest in "single-frame estimator-aided attitude filtering" is higher in the more recent literature [9–11]. The attitude determination concept of the Kyushu University mini-satellite QSAT is based on a combination of the *Weighted-Least-Square* (WLS) and KF [9, 10]. The WLS method produces the optimal attitude-angle observations at a single-frame by using the Sun sensor and magnetometer measurements. The KF combines the WLS angular observations with the attitude rate measured by gyros to produce the optimal attitude solution. In [11], an interlaced filtering method is presented for determination of the nanosatellite attitude. In this integrated system, the optimal-REQUEST and UKF algorithms are combined to estimate the attitude quaternion and gyro drifts. The optimal-REQUEST, which cannot estimate gyroscope drifts, is run for the attitude estimation. Then, the UKF is used for the gyro-drift estimation on the basis of linear measurement results obtained as the optimal-REQUEST estimates. There are also similar applications for the UAV attitude estimation. De Marina et al. introduce an attitude heading reference system based on the UKF using the TRIAD algorithm as the observation model in [12].

Here, we may also refer to the studies where a single-frame attitude estimator is used together with an attitude filter but does not provide linear measurements [13, 14]. For linear measurements, it is equivalent to first updating the attitude using the single-frame estimator and subsequently using this updated portion of the state to updating the remainder of the state as if updating the entire state at once. However in [13, 14], the measurement model is a nonlinear one. A nonlinear updating the attitude is obtained by solving the Wahba's problem and subsequently updating the non-attitude states using the optimal gain for the linear measurement case. Therefore, in these studies, the attitude is updated using a single-frame estimator, whereas all remaining non-attitude states are updated using standard nonlinear attitude filters.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 113–125.

In this study we examine an *SVD-aided UKF* (SaUKF) algorithm for the nanosatellite attitude estimation. The nanosatellite has magnetometers and Sun sensors as on-board attitude sensors. In the first phase, Wahba's problem is solved by the *Singular Value Decomposition* (SVD) method and quaternion estimations are obtained for the satellite's attitude. These quaternion estimations are then used as the measurement results for an UKF, which forms the second phase of the algorithm. The SaUKF provides improved attitude knowledge and attitude rate estimates. The whole algorithm runs recursively. The main aim is to propose an easy-to-apply and accurate nanosatellite attitude estimation algorithm, which is also robust against estimation deteriorations when the satellite is in its eclipse. The initial results are presented in [15]. In this study we compare the results with those obtained by an UKF that uses nonlinear measurements. Besides we propose an algorithm that switches between the UKF with nonlinear measurements and the SaUKF to ensure both the accuracy and robustness.

## 2. Satellite equations of motion and measurement models

In this section we briefly review the satellite equations of motion and the measurement models for magnetometers and Sun sensors.

### *2.1. Satellite equations of motion*

The satellite's kinematics equation of motion derived using the quaternion attitude representation can be presented as [16]:

$$\dot{\boldsymbol{q}}(t) = \frac{1}{2}\Omega(\boldsymbol{\omega}_{BR}(t))\boldsymbol{q}(t). \tag{1}$$

In (1), the quaternion $\boldsymbol{q}$ is composed of four attitude parameters, $\boldsymbol{q} = \begin{bmatrix} q_1 & q_2 & q_3 & q_4 \end{bmatrix}^T$. Three terms of the quaternion $\boldsymbol{q}$ are vectors, whereas the last term is a scalar. Then, the quaternion can take a form of $\boldsymbol{q} = \begin{bmatrix} \boldsymbol{g}^T & q_4 \end{bmatrix}^T$, $\boldsymbol{g} = \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix}^T$. Moreover, in (1), $\Omega(\boldsymbol{\omega}_{BR})$ is the skew symmetric matrix as:

$$\Omega(\boldsymbol{\omega}_{BR}) = \begin{bmatrix} 0 & \omega_3 & -\omega_2 & \omega_1 \\ -\omega_3 & 0 & \omega_1 & \omega_2 \\ \omega_2 & -\omega_1 & 0 & \omega_3 \\ -\omega_1 & -\omega_2 & -\omega_3 & 0 \end{bmatrix}, \tag{2}$$

where the $\boldsymbol{\omega}_{BR}$ vector is composed of $\omega_1$, $\omega_2$ and $\omega_3$; it indicates the angular velocity of the body frame with respect to the orbit frame. The angular rate vector should be identified because of the sensor usage. Hence, the rate vector in the body frame with respect to the inertial coordinate system can be shown as: $\boldsymbol{\omega}_{BI} = \begin{bmatrix} \omega_x & \omega_y & \omega_z \end{bmatrix}^T$. $\boldsymbol{\omega}_{BI}$ and $\boldsymbol{\omega}_{BR}$ can be related according to the following equation:

$$\boldsymbol{\omega}_{BR} = \boldsymbol{\omega}_{BI} - A\begin{bmatrix} 0 & -\omega_o & 0 \end{bmatrix}^T. \tag{3}$$

The angular velocity of the satellite on its orbit is specified by $\omega_o$ with respect to the inertial reference, found as $\omega_o = \left(\mu / r^3\right)^{1/2}$ for a circular orbit using $\mu$, which is the product of two constants $\left(GM_E\right)$. Here, $G$ is the gravitational constant, $M_E$ – the mass of the Earth and $r$ – the distance between the satellite and Earth centers of masses. In (3) $A$ is a transformation matrix which can be related to the quaternions as follows:

$$A = (q_4^2 - \boldsymbol{g}^2) I_{3\times 3} + 2\boldsymbol{g}\boldsymbol{g}^T - 2q_4 [\boldsymbol{g}\times]. \tag{4}$$

The unit matrix $I_{3x3}$ has a dimension of $3\times 3$ and $[\boldsymbol{g}\times]$ is a skew-symmetric matrix as follows:

$$[\boldsymbol{g}\times] = \begin{bmatrix} 0 & -g_3 & g_2 \\ g_3 & 0 & -g_1 \\ -g_2 & g_1 & 0 \end{bmatrix}. \tag{5}$$

The satellite's dynamic equations are necessary to estimate the full state attitude including both the attitude and attitude rates. Based on the Euler's equations the dynamic knowledge can be found by:

$$J\frac{\boldsymbol{\omega}_{BI}}{dt} = \boldsymbol{N}_d - \boldsymbol{\omega}_{BI} \times (J\boldsymbol{\omega}_{BI}), \tag{6}$$

where $J$ is an inertia matrix composed of $J = diag(J_x, J_y, J_z)$ which are the principal moments of inertia. The external torques affecting the satellite can be added in order to find the resulting disturbance torque, $\boldsymbol{N}_d$:

$$\boldsymbol{N}_d = \boldsymbol{N}_{gg} + \boldsymbol{N}_{ad} + \boldsymbol{N}_{sp} + \boldsymbol{N}_{md}, \tag{7}$$

where $\boldsymbol{N}_{gg}$ is the gravity gradient torque, $\boldsymbol{N}_{ad}$ is the aerodynamic disturbance torque, $\boldsymbol{N}_{sp}$ is the solar pressure disturbance torque and $\boldsymbol{N}_{md}$ is the residual magnetic torque caused by the interaction of the satellite's residual dipole and the Earth's magnetic field [16].

### *2.2. Sensor models*

The magnetometer sensor for attitude determination is a very common sensor for small satellite missions. A model of the Earth's magnetic field measurements can be given in (8) (the magnetometers are assumed to be calibrated) [17, 18]:

$$\begin{bmatrix} B_x(\boldsymbol{q},t) \\ B_y(\boldsymbol{q},t) \\ B_z(\boldsymbol{q},t) \end{bmatrix} = A \begin{bmatrix} B_1(t) \\ B_2(t) \\ B_3(t) \end{bmatrix} + \eta_1. \tag{8}$$

The components of the Earth's magnetic field, $B_1(t), B_2(t)$ and $B_3(t)$, in the orbital coordinate frame can be calculated by the common and accurate magnetic field model, *International Geomagnetic Reference Field* (IGRF) [19]. $B_x(\boldsymbol{q},t), B_y(\boldsymbol{q},t)$ and $B_z(\boldsymbol{q},t)$ are the vector components of magnetic field measured by the magnetometers. Therefore, they are presented in the body reference system. Moreover, $\eta_1$ is the zero mean Gaussian white noise:

$$E[\eta_{1k}\eta_{1j}^T] = I_{3x3}\sigma_m^2\delta_{kj}, \tag{9}$$

where $\sigma_m$ is the standard deviation and $\delta_{kj}$ is the Kronecker symbol.

The Sun direction with respect to the inertial coordinates regarding the Earth center depends only on time referred to Julian Day ($T_{TDB}$). $T_{TDB}$ can be derived using the satellite's reference epoch and the exact time. The variables are the mean anomaly ($M_{Sun}$) and the mean longitude ($\lambda_{M_{Sun}}$) of the Sun. Using (10), the ecliptic longitude of the Sun ($\lambda_{ecliptic}$) and its linear model ($\varepsilon$) can be found [20]:

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 113–125.

$$M_{Sun} = 357.5277233^0 + 35999.05034T_{TDB}\,, \tag{10a}$$

$$\lambda_{ecliptic} = \lambda_{M_{Sun}} + 1.914666471^0 sin(M_{Sun}) + 0.019994643 sin(2M_{Sun})\,, \tag{10b}$$

$$\lambda_{M_{Sun}} = 280.4606184^\circ + 36000.77005361T_{TDB}\,, \tag{10c}$$

$$\varepsilon = 23.439291^0 - 0.0130042T_{TDB}\,. \tag{10d}$$

From those relations (10), the Sun direction vector ($\boldsymbol{S}_{ECI}$) in the inertial coordinates can be found:

$$\boldsymbol{S}_{ECI} = \begin{bmatrix} \cos\lambda_{ecliptic} \\ \sin\lambda_{ecliptic}\cos\varepsilon \\ \sin\lambda_{ecliptic}\sin\varepsilon \end{bmatrix}. \tag{11}$$

However, since the satellite is rotating along its trajectory, it is necessary to transform the unit Sun direction vector into the orbital frame by using the orbit propagation algorithm. Finally, the (12) shows the relation between the Sun sensor measurement vector and the Sun direction model vector:

$$\boldsymbol{S}_b = A\boldsymbol{S}_o + \eta_2\,, \tag{12}$$

where $\boldsymbol{S}_o$ is the Sun direction vector in the orbit reference system and $\boldsymbol{S}_b$ is the vector of Sun sensor measurements in the body reference system having the zero mean Gaussian white noise $\eta_2$ with the characteristic of:

$$E\left[\eta_{2k}\eta_{2j}^T\right] = I_{3x3}\sigma_s^2\delta_{kj}\,, \tag{13}$$

where $\sigma_s$ is the standard deviation of Sun sensor error.

The satellite's orbital elements and its position on the orbit must be known to model the Earth's magnetic field and Sun vectors in the orbit frame.

## 3. SVD-aided UKF algorithm

The contents of this section include estimation of the satellite's attitude and the angular velocities during the operational mode of the mission. The estimation process is divided into two stages: SVD and UKF. Firstly, a single frame method SVD minimizes the Wahba's loss function by using two vectors and finds the coarse attitude angles and variance values for each axis. Then, UKF uses the SVD results as the input values in each time step and provides the filtered attitude and attitude rates with a higher accuracy.

### *3.1. SVD method*

As a single-frame method, SVD aims to solve the problem formulated by Grace Wahba [21]. In every single time frame SVD can estimate the coarse attitude only by using the measurement results and the model vectors. In the loss function (see (14)), $\mathbf{b}_i$ and $\mathbf{r}_i$ are sets of unit vectors obtained in two different coordinate systems in every single time interval. From the optimal solution for the orthogonal $A$ matrix, the attitude angles can be found [22]:

$$L(A) = \frac{1}{2}\sum_i a_i \left|\mathbf{b}_i - A\mathbf{r}_i\right|^2. \tag{14}$$

The unit vectors in the loss function represent the Sun direction and Earth's magnetic field vectors for the orbit frame ($\mathbf{r}_i$) and the body frame ($\mathbf{b}_i$), where $a_i$ is a non-negative weight. The loss can be reduced in (15) as:

$$L(A) = \lambda_0 - \mathrm{tr}\left(AB^T\right), \tag{15}$$

where:

$$\lambda_0 = \sum a_i, \tag{16a}$$

$$B = \sum a_i \mathbf{b}_i \mathbf{r}_i^T. \tag{16b}$$

The SVD method can be used here to maximize the trace function expressed in the (15) by using the most robust algorithm from single-frame methods [22]. $B$ matrix has the singular value decomposition:

$$B = U \sum^T V^T = U\mathrm{diag}\left[\sum_{11} \quad \sum_{22} \quad \sum_{33}\right] V^T, \tag{17}$$

where matrices $U$ and $V$ are orthogonal and the singular values hold $\sum_{11} \geq \sum_{22} \geq \sum_{33} \geq 0$. Then, the optimal attitude matrix can be found:

$$U^T A_{opt} V = \mathrm{diag}[1 \quad 1 \quad \det(U)\det(V)], \tag{18}$$

$$A_{opt} = U\mathrm{diag}[1 \quad 1 \quad \det(U)\det(V)]V^T. \tag{19}$$

The covariance analysis is an important process in the integrated filtering technique and the matrix $P_{svd}$ can be obtained by defining secondary singular values $s_1 = \sum_{11}$, $s_2 = \sum_{22}$, $s_3 = det(U)det(V)\sum_{33}$, as follows:

$$P_{svd} = U\mathrm{diag}[(\mathrm{s}_2 + \mathrm{s}_3)^{-1} \quad (\mathrm{s}_3 + \mathrm{s}_1)^{-1} \quad (\mathrm{s}_1 + \mathrm{s}_2)^{-1}]U^T. \tag{20}$$

The method requires measurements at every single moment to accurately provide the attitude angles. Hence, the method fails when either the satellite is in the eclipse or two vectors are parallel.

### *3.2. Unscented Kalman Filter*

The UKF uses an accurate approximation called the Unscented Transform for solving the multidimensional integrals instead of the linear approximation to the nonlinear equations as *Extended Kalman Filter* (EKF) does [23]. The essence is the fact that the approximation of a nonlinear distribution is easier than the approximation of a nonlinear function or transformation. The conventional algorithm for the UKF is not presented here for brevity and the reader may refer to [24], specifically for attitude estimation using the UKF.

When a quaternion in the kinematic modeling of the satellite's motion is used, the UKF in a standard format cannot be implemented straightforwardly. The reason of such a drawback is the constraint of quaternion unity expressed by $\boldsymbol{q}^T\boldsymbol{q} = 1$. If the kinematics (1) is used in the filter directly, than there is no guarantee that the predicted quaternion mean of the UKF will satisfy this constraint.

In the reference [24], the authors overcome this problem by using an unconstrained three-component vector to represent an attitude-error quaternion instead of using all four components of the quaternion vector. They represent the local error-quaternion by the vector of *Generalized Rodrigues Parameters* (GRP). In this paper we use the same method.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 113–125.

Recall that we represent a quaternion with its vector and scalar parts as $\boldsymbol{q} = \begin{bmatrix} \boldsymbol{g}^T & q_4 \end{bmatrix}^T$. After that, when the local error-quaternion is denoted by $\delta\boldsymbol{q} = \begin{bmatrix} \delta\boldsymbol{g}^T & \delta q_4 \end{bmatrix}^T$, the vector of GRP may be given as:

$$\delta\boldsymbol{p} = f\left[\delta\boldsymbol{g} / (a + \delta q_4)\right], \tag{21}$$

where $a$ is a parameter from 0 to1 and $f$ is the scale factor. When $a = 0$ and $f = 1$ then (21) gives the Gibbs vector, whereas when $a = 1$ and $f = 1$ then (21) gives the standard vector of modified Rodrigues parameters. In the paper [24] – as well as in this paper – $f$ is chosen as $f = 2(a+1)$. The inverse transformation from $\delta\boldsymbol{p}$ to $\delta\boldsymbol{q}$ is given by:

$$\delta q_4 = \frac{-a\left\|\delta\boldsymbol{p}\right\|^2 + f\sqrt{f^2 + (1-a^2)\left\|\delta\boldsymbol{p}\right\|^2}}{f^2 + \left\|\delta\boldsymbol{p}\right\|^2}, \tag{22a}$$

$$\delta\boldsymbol{g} = f^{-1}(a + \delta q_4)\delta\boldsymbol{p}. \tag{22b}$$

### *3.3. Estimation of attitude and attitude rate using SaUKF*

Two methods are integrated and the SaUKF algorithm is proposed for the nanosatellite attitude estimation. The main purposes are:

1. As a standalone technique the SVD works well as long as minimum 2 vector measurements are available and not parallel. However, if there is only one vector measurement when the satellite is in the eclipse, the SVD fails to provide any attitude estimate.
2. The SVD method gives attitude estimates as frequent as the sampling rate of the sensor with a lower measurement frequency (if there is no propagation). The SaUKF can provide the attitude estimate with a higher frequency since it makes use of the attitude dynamics.
3. The SVD method does not estimate attitude rates. For most of the cases the satellite attitude rates have to be estimated – especially for control purposes. There are deterministic methods to estimate the satellite's attitude rate from the vector measurement results [25], but usually a filtering-based method gives more accurate estimates.

When the SVD method cannot give any estimation results, the covariance for the SVD estimations – and so the elements of the $R$ matrix – increase. Therefore, the UKF is robust against the failures in the SVD estimations, as we see during the eclipse period.

As the attitude representation, in SVD algorithm there are used quaternions. However, for the SaUKF, the attitude errors regarding GRP are acquired:

$$\delta\boldsymbol{q}_{obs} = \boldsymbol{q}_{mes} \otimes \left[\hat{\boldsymbol{q}}_0(k+1|k)\right]^{-1}, \tag{23}$$

where $\boldsymbol{q}_{mes}$, coming from the SVD method, are quaternion-multiplied with the predicted mean quaternion. Then, regarding $\delta\boldsymbol{q}_{obs} = \begin{bmatrix} \delta\boldsymbol{g}_{obs}^T & \delta q_{4,obs} \end{bmatrix}^T$, the measurement result of the attitude error is calculated as:

$$\delta\boldsymbol{p}_{obs} = f\left[\delta\boldsymbol{g}_{obs} / (a + \delta q_{4,obs})\right]. \tag{24}$$

A scheme of the attitude and rate estimation algorithm of the integrated method is given in Fig. 1.

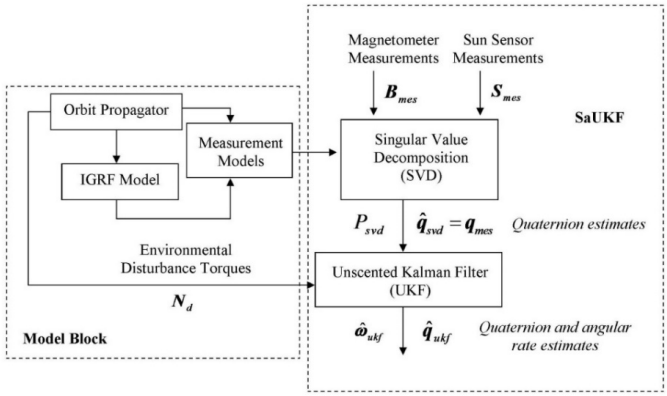



Fig. 1. A scheme of the attitude and attitude rate estimation using the SaUKF.

## 4. Simulations for nanosatellite

Several simulations were performed in order to evaluate the attitude estimation algorithm. A three-unit cube-sized satellite with about 3 kg mass and $J = diag(0.055\ 0.055\ 0.017)$ kg.m$^2$ inertia matrix is considered for the estimation scheme. The satellite has an almost circular orbit with an eccentricity of $e = 6.4 \times 10^{-5}$ and $i = 74°$ inclination at 612 km altitude.

All sensors are assumed to be calibrated against biases, scale factors and so on. Therefore, only the sensor noise (zero mean Gaussian white noise) is considered in the algorithm with $\sigma_m = 300nT$ standard deviation for the magnetometer and $\sigma_s = 0.002$ unit for the Sun sensor. The total orbital time is close to 6000 sec. and the time step is taken as 1 sec.

Both SaUKF and UKF use the process noise covariance values of $1\times10^{-4}$ and $1\times10^{-9}$ for attitude and rates, respectively, and have an eclipse period between 2000–4000 sec. In Fig. 2, the estimation error results for SaUKF, SVD and UKF only can be seen and compared. It is clearly seen that the SaUKF estimates the attitude more accurately than both the SVD and UKF only methods, with the exception of the eclipse period. During the eclipse period the SVD method fails because no Sun sensor data are obtained. The quaternion measurements for the SaUKF deteriorate and the values of *R*, which are coming from the covariance matrix of SVD angle estimation errors ($P_{svd}$), increase. If the SaUKF gain values become very low (since *R* values are very high), the correction term of the UKF will become insignificant and the contribution of the propagation model to estimation becomes dominant. That enables the attitude estimation during the eclipse period, even though there is no measurement input to the filter. As it is seen in Fig. 2, the proposed SaUKF method convergences slower than the traditional UKF. This is a drawback of the presented SaUKF method. Therefore, it is recommended to use the proposed method after the convergence of the nontraditional UKF.

The process noise covariance Q is a parameter that enables the filter to base mostly on either the measurements or the dynamics in the filter. In the filter, $1\times10^{-4}$ and $1\times10^{-9}$ pair is used as medium noise. Here, at the end of the eclipse period, before the Sun sensor data arrival, the attitude angle has an error of 10 degrees. If the Q pair is $1\times10^{-3}$ and $1\times10^{-7}$, which is higher than the selected one, the results are close to the measurement ones and the attitude angles are diverging more during the eclipse period. On the other hand, for lower pair values – such as $1\times10^{-9}$ and $1\times10^{-13}$ – the SaUKF becomes non-agile, i.e. has a smaller convergence rate at the end of the eclipse or the beginning of the orbit (Fig. 3).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 113–125.

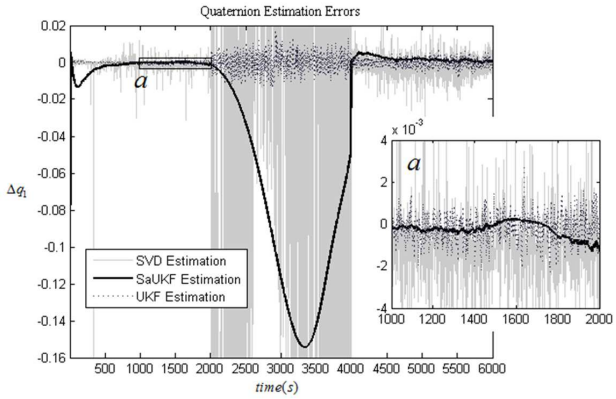



Fig. 2. The estimation error for quaternion q1; comparison of the UKF and SVD only estimations with those of the SaUKF. The subfigure a zooms to the area indicated in the main figure.

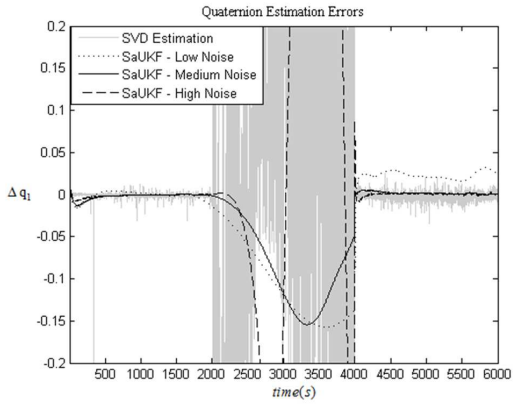



Fig. 3. The quaternion estimation error for the SaUKF with different values of process noise covariance Q.

In the eclipse period, the UKF only method gives the most accurate attitude estimations. During that period it works only with the magnetometer measurements. Since the magnetometers are coarser sensors comparing with the Sun sensors, there is a clear increase in the UKF estimation error in the eclipse but still the estimations are accurate enough for a nanosatellite mission with this sensor configuration (less than 0.1 degrees – see Fig. 4 for the attitude estimation error norms).

The angular velocities of the satellite for each axis can be estimated accurately by using the SaUKF (see Fig. 5). During the eclipse period the attitude rate estimations are not deteriorated as much as the attitude estimates resulting from accurate dynamic knowledge and low process noise for dynamics propagation. The rate estimates obtained by the UKF are similar.

The main disadvantage of the proposed SaUKF method is the requirement of accurate measurements – free of any bias, sensor misalignment and other sorts of errors. The sensors must be calibrated before using their measurement results as an input to the SaUKF. As discussed in several papers [1, 2, 18] particularly for the magnetometers, such a calibration should be performed on-orbit for nanosatellite missions. In addition, as it is clearly demonstrated by the simulation results, the estimation performance of the SaUKF degrades during the eclipse period and the UKF based on the results of nonlinear measurements provides more accurate estimations. Regarding these facts, our suggestion is to use an algorithm which

switches between several different filters in accordance with the flight mode. An example is given in Fig. 6.

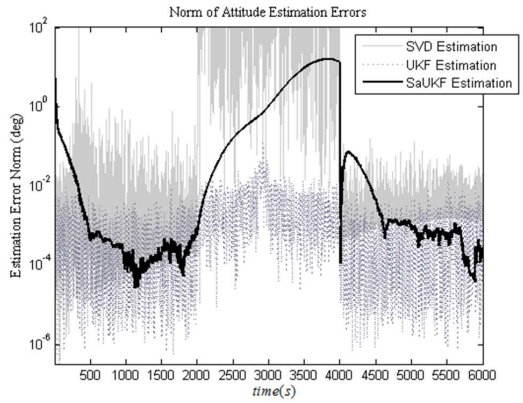



Fig. 4. The norm of attitude estimation errors.

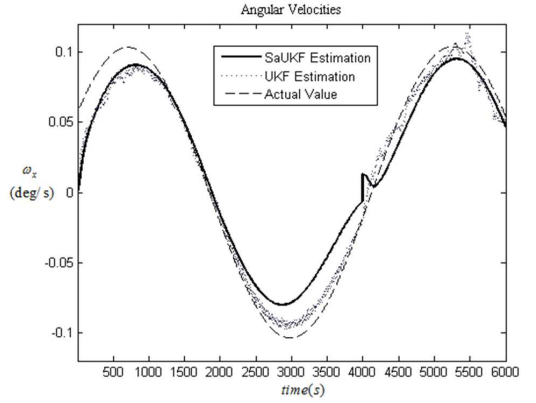



Fig. 5. Estimation of the angular rate along the x axis.

In Fig. 7, two methods are switched in/out of the eclipse period for the more accurate attitude estimation. As mentioned earlier, the SaUKF estimates the attitude more accurately than both SVD and UKF only methods except for the eclipse period; that is why the SaUKF algorithm is used only outside the eclipse. When the satellite is on the dark side of the Earth, the SVD method fails since it is fed with no Sun sensor measurements. The results of the UKF only method are presented in Fig. 7. Also, it should be kept in mind that the switching between the algorithms should be managed after the stabilization of the satellite because right after the eclipse period tumbling may occur.

Certainly, for the nanosatellite application we also need to examine the computational load of each algorithm. Table.1 gives the running times of the algorithms for 6000 sec. simulation, details of which are discussed above. The simulations are performed on a computer with Intel® Core™ i7 @2.93 GHz CPU and 3.49 GB RAM. It shall be noted that all the presented data include the computation time required for simulating the real attitude and measurements. We see that, for the SaUKF algorithm, the SVD is the computationally heavier part and the SaUKF requires a higher load comparing with the UKF based on nonlinear measurements. Yet, the load is not so heavy as to prevent a nanosatellite application, especially if we consider the recent improvements in microprocessors capacity.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 113–125.

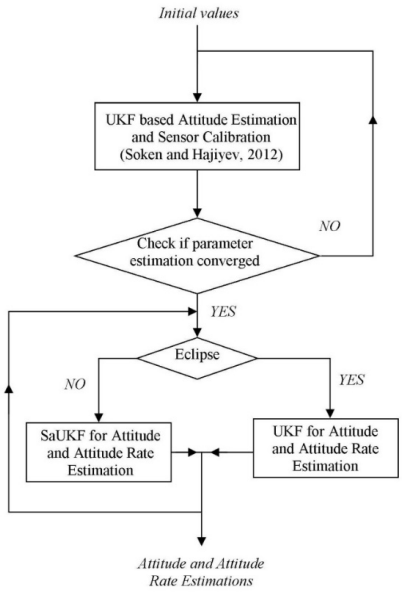



Fig. 6. A block diagram of attitude and attitude rate estimation for the proposed algorithm.

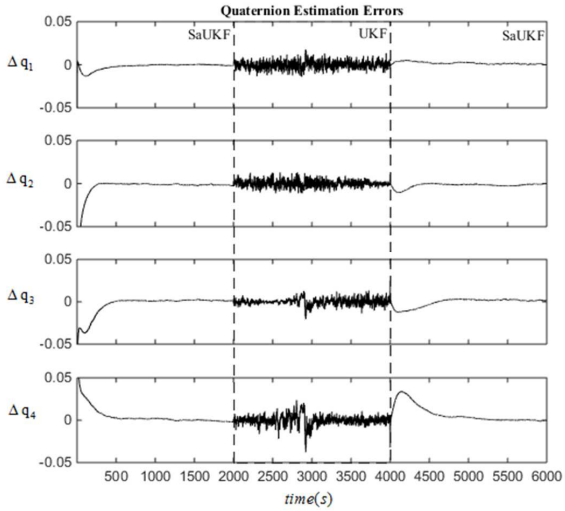



Fig. 7. Estimation of the quaternions by the SaUKF (outside the eclipse period) and UKF (in the eclipse period).

Table 1. The computation times for each algorithm.

| Computation time (sec) for 10 Monte Carlo runs | SVD | SaUKF | UKF |
|---|---|---|---|
| | 14.30 | 17.96 | 10.49 |

## 5. Conclusion

In this paper, the *Singular Value Decomposition* (SVD) method and *Unscented Kalman Filter* (UKF) are integrated to determine the attitude and attitude rate for a three-unit cube-

sized satellite. The quaternion representation is used to avoid any singularities based on the trigonometric equations. The SVD method fails in the eclipse period because of no Sun observation results. On the other hand, the *SVD-aided UKF* (SaUKF) can estimate the attitude even in the eclipse period, although it is a coarse estimate. The simulation results show that also the UKF with nonlinear vector measurements ensures a reasonable accuracy of the attitude estimation. In the eclipse period the accuracy of the UKF is higher than that of the SaUKF; beyond that period the SaUKF is the most accurate estimation method.

The simulation results show that the proposed SaUKF method convergences slower than the traditional UKF. This is a drawback of the SaUKF method. Therefore, it is recommended to use the proposed method after the convergence of a nontraditional UKF. The ideal algorithm that we suggest for the examined case is composed of the SaUKF and UKF. The SaUKF is used when the Sun sensor measurements are available; in the eclipse period the algorithm switches to the UKF.

## Acknowledgments

The work was supported by TUBITAK (The Scientific and Technological Research Council of Turkey), Grant 113E595.

## References

[1] Vinther, K., Jensen, K.F., Larsen, J.A., Wisniewski, R. (2011). Inexpensive cubesat attitude estimation using quaternions and unscented Kalman filtering. *Automatic Control in Aerospace*, 4(1).

[2] Springmann, J.C., Cutler, J.W. (2014). Flight results of a low-cost attitude determination systems. *Acta Astronautica*, 99, 201–204.

[3] Markley, F.L. (1999). Attitude determination using two vector measurements. *Flight Mechanics Symposium*, 39–52, Goddard Space Flight Center, Greenbelt, MD.

[4] Cordova-Alarcon, J.R., Mendoza-Barcenas, M.A., Solis-Santome, A. (2015). Attitude Determination System Based on Vector Observations for Satellites Experiencing Sun-Eclipse Phases. *Multibody Mechatronic Systems*, Springer International Publishing, 75–85.

[5] Psiaki, M.L. (1989). Three-axis attitude determination via Kalman filtering of magnetometer data. *Journal of Guidance, Control and Dynamics*, 13(3), 506–514.

[6] Sekhavat, P., Gong, Q., Ross, I.M. (2007). NPSAT I parameter estimation using unscented Kalman filter. *Proc. 2007 American Control Conference*, New York, USA, 4445–451.

[7] Cilden, D., Hajiyev, C. (2015). Error Analysis of the Vector Measurements Based Attitude Determination Methods for Small Satellites. *International Symposium on Space Technology and Science (ISTS)*, Kobe, Hyogo, Japan.

[8] Hajiyev, C., Bahar, M. (2003). Attitude determination and control system design of the ITU-UUBF LEO1 satellite. *Acta Astronautica*, 52(2–6), 493–499.

[9] Mimasu, B.Y., Van der Ha, J.C., Narumi, T. (2008). Attitude determination by magnetometer and gyros during eclipse. *AIAA/AAS Astrodynamics Specialist Conference and Exhibit*, Honolulu, USA.

[10] Mimasu, B.Y., Van der Ha, J.C. (2009). Attitude determination concept for QSAT. *Transactions of the Japan Society for Aeronautical and Space Sciences, Aerospace Technology Japan*, 7, 63–68.

[11] Quan, W., Xu, L., Zhang, H., Fang, J. (2013). Interlaced Optimal-REQUEST and unscented Kalman filtering for attitude determination. *Chinese Journal of Aeronautics*, 26(2), 449–155.

[12] de Marina, H.G., Espinosa, F., Santos, C. (2012). Adaptive UAV Attitude Estimation Employing Unscented Kalman Filter, FOAM and Low-Cost MEMS Sensors. *Sensors*, 12(7), 9566–9585.

[13] Christian, J.A., Lightsey, E.G. (2010). Sequential optimal attitude recursion filter. *Journal of Guidance, Control, and Dynamics*, 33(6), 1787–1800.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 113–125.

[14] Ainscough, T., Zanetti, R. (2014). Q-Method extended Kalman filter. *Journal of Guidance, Control, and Dynamics*.

[15] Cilden, D., Hajiyev, C., Soken, H.E. (2015). Attitude and Attitude Rate Estimation for a Nanosatellite Using SVD and UKF. *Recent Advances in Space Technologies*, Istanbul, Turkey.

[16] Wertz, J.R. (1988). *Spacecraft Attitude Determination and Control*. Dordrecht, Holland: Kluwer Academic Publishers.

[17] Alonso, R., Shuster, M.D. (2002). Complete linear attitude-independent magnetometer calibration. *Journal of Astronautical Sciences*, 50, 477–490.

[18] Soken, H.E., Hajiyev, C. (2012). UKF-Based reconfigurable attitude parameters estimation and magnetometer calibration. *IEEE Transactions on Aerospace and Electronic Systems*, 48(3), 2614–2627.

[19] Finlay, C., Maus, S., Beggan, C.D., Bondar, T.N., *et al.* (2010). International Geomagnetic Reference Field: the eleventh generation. *Geophysical Journal International*, 183, 1216–1230.

[20] Vallado, D.A. (2007). *Fundamentals of Astrodynamics and Applications.* Space Technology Library. USA: Microcosm Press/Springer, 21.

[21] Wahba, G. (1965). Problem 65–1: A Least Squares Estimate of Satellite Attitude. *Society for Industrial and Applied Mathematics Review*, 7(3), 409.

[22] Markley, F.L., Mortari, D. (2000). Quaternion attitude estimation using vector observations. *Journal of the Astronautical Sciences*, 48(2–3), 359–380.

[23] Julier, S.J., Uhlmann, J.K., Durrant-Whyte, H.F. (1995). A new approach for filtering nonlinear systems. *American Control Conference*, Seattle, USA. 1628–1632.

[24] Crassidis, J.L., Markley, F.L. (2003). Unscented filtering for spacecraft attitude estimation. *Journal of Guidance Control and Dynamics*, 26(4), 536–542.

[25] Oshman, Y., Dellus, F. (2003). Spacecraft angular velocity estimation using sequential observations of a single directional vector. *Journal of Spacecraft and Rockets*, 40(2), 234–247.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 127–142.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# ESTIMATION OF UAV POSITION WITH USE OF SMOOTHING ALGORITHMS

**Piotr Kaniewski, Rafał Gil, Stanisław Konatowski**

*Military University of Technology, Institute of Radioelectronics, Gen. S. Kaliski 2, 00-908 Warsaw, Poland*
*(✉ pkaniewski@wat.edu.pl, +48 261 839 080, rgil@wat.edu.pl, skonatowski@wat.edu.pl)*

**Abstract**

The paper presents methods of on-line and off-line estimation of UAV position on the basis of measurements from its integrated navigation system. The navigation system installed on board UAV contains an INS and a GNSS receiver. The UAV position, as well as its velocity and orientation are estimated with the use of smoothing algorithms. For off-line estimation, a fixed-interval smoothing algorithm has been applied. On-line estimation has been accomplished with the use of a fixed-lag smoothing algorithm. The paper includes chosen results of simulations demonstrating improvements of accuracy of UAV position estimation with the use of smoothing algorithms in comparison with the use of a Kalman filter.

Keywords: Unmanned Aerial Vehicle, Inertial Navigation System, Global Navigation Satellite System, Integrated Navigation System, Synthetic Aperture Radar, Kalman Filter, Smoothing Algorithm.



## 1. Introduction

*Unmanned Aerial* Vehicles (UAV) are becoming more and more popular in military and civilian applications. In military context, they are applied mainly in *Electronic Intelligence* (ELINT) and *Imagery Intelligence* (IMINT) [1], which includes radar terrain imaging with the use of *Synthetic Aperture Radar* (SAR) [2–4].

The autonomy of operation requires that the position, velocity and angular orientation of UAV are estimated on-line and used for appropriate execution of its mission. Such an estimation is usually accomplished in an on-board integrated navigation system, typically composed of an *Inertial Navigation System* (INS) [5, 6] and a *Global Navigation Satellite System* (GNSS) receiver [5, 7], with the use of some form of *Kalman Filter* (KF) [8–15].

In some applications, *e.g.* SAR imagery, requirements with respect to the accuracy of positioning are very high [3, 4, 16, 17]. On the other hand, short delays in image availability are often acceptable. Thus, in this group of applications, fixed-lag smoothing algorithms [18, 19], which provide delayed but more accurate estimates than Kalman filters, can be applied to on-line estimation of UAV position. There are also applications where UAV trajectory and parameters of flight can be reconstructed off-line, after mission, on the basis of logged navigation data. In such a case, the authors suggest using very accurate fixed-interval smoothing algorithms [10, 18].

The layout of the paper is as follows. Firstly, a state-space model of an integrated navigation system used on-board UAV is presented. It is assumed that the system is loosely integrated according to the compensation method with feed-forward correction [5, 19] and is composed of an INS and a GNSS receiver. Such an INS/GNSS system has been designed and produced within the scope of the WATSAR project, performed by the Military University of Technology, Warsaw, Poland, and a Polish private company WB Electronics S.A. [17, 20]. Subsequently, the fixed-interval and the fixed-lag smoothing algorithms are described. The paper includes also chosen results of simulations, demonstrating improvements of accuracy of UAV position

Article history: received on Jul. 28, 2016; accepted on Oct. 09, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0013.

estimation with the use of smoothing algorithms in comparison with the one using a Kalman filter. Finally, a discussion of the results and conclusions are presented.

## 2. State-space model of INS/GNSS system

Implementation of a Kalman filter or a smoothing algorithm in an INS/GNSS system requires previous formulation of its state-space model [5, 7]. In the case of a loosely integrated system [9], designed in the WATSAR project [17, 20], the discrete state-space model is linear and it is given by a pair of equations [11, 13–15]:

$$\mathbf{x}(k+1) = \mathbf{\Phi}(k+1,k)\mathbf{x}(k) + \mathbf{w}(k), \tag{1}$$

$$\mathbf{z}(k+1) = \mathbf{H}(k+1)\mathbf{x}(k+1) + \mathbf{v}(k+1), \tag{2}$$

where: $\mathbf{x}$ – a state vector; $\mathbf{w}$ – a vector of discrete random process disturbances; $\mathbf{z}$ – a measurement vector; $\mathbf{v}$ – a vector of measurement errors; $\mathbf{\Phi}$ – a transition matrix; $\mathbf{H}$ – an observation matrix.

Equation (1) is called the dynamics model and for the designed INS/GNSS system it describes propagation in time of errors of a custom-built INS. These errors include position, velocity and orientation errors resulting from processing erroneous inertial data inside the INS. Detailed INS errors models can be very complicated and may contain even several tens of states [7]. Some of these states are observable only conditionally, *e.g.* during maneuvers of UAV, and only in high-quality navigation-grade inertial systems. As in the WATSAR project only a medium-quality, tactical-grade INS has been used, a simple 9-state model of INS errors has been applied [17], with 3 states for position errors, 3 states for velocity errors and 3 states for orientation errors with respect to various axes of the local reference horizontal system of coordinates NED (North-East-Down) [9].

The 9-state dynamics model is originally continuous and it is based on a set of 9 scalar first-order differential equations, describing the relationship between the states constituting the state vector $\mathbf{x}$ and their first derivatives [5, 7]:

$$\delta\dot{N} = \delta v_N, \tag{3}$$

$$\delta\dot{v}_N = -f_D\phi_E + f_E\phi_D + u_{vN}, \tag{4}$$

$$\dot{\phi}_E = -\frac{1}{R}\delta v_N + \omega_N\phi_D + u_{\phi E}, \tag{5}$$

$$\delta\dot{E} = \delta v_E, \tag{6}$$

$$\delta\dot{v}_E = f_D\phi_N - f_N\phi_D + u_{vE}, \tag{7}$$

$$\dot{\phi}_N = \frac{1}{R}\delta v_E - \omega_E\phi_D + u_{\phi N}, \tag{8}$$

$$\delta\dot{D} = \delta v_D, \tag{9}$$

$$\delta\dot{v}_D = \frac{2g}{R}\delta D + u_{vD}, \tag{10}$$

$$\dot{\phi}_D = u_{\phi D}, \tag{11}$$

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 127–142.

where: $\delta N$, $\delta E$, $\delta D$ – INS position errors along the North, East and Down axes; $\delta v_N$, $\delta v_E$, $\delta v_D$ – INS velocity errors along the North, East and Down axes; $\phi_N$, $\phi_E$, $\phi_D$ – INS attitude errors around the North, East and Down axes; $f_N$, $f_E$, $f_D$ – specific forces along the North, East and Down axes; $\omega_N$, $\omega_E$ – components of the angular velocity around the North and East axes; $g$ – gravity acceleration; $R$ – the Earth's radius in the spherical model; $u_{vN}$, $u_{vE}$, $u_{vD}$ – errors of INS accelerometers; $u_{\phi N}$, $u_{\phi E}$, $u_{\phi D}$ – errors of INS gyros.

Grouping the above set of scalar equations into a single equation, we obtain the following continuous dynamics model of the system:

$$\underbrace{\frac{d}{dt}\begin{bmatrix}\delta N\\ \delta v_N\\ \phi_E\\ \delta E\\ \delta v_E\\ \phi_N\\ \delta D\\ \delta v_D\\ \phi_D\end{bmatrix}}_{\dot{\mathbf{x}}(t)} = \underbrace{\begin{bmatrix}0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & -f_D & 0 & 0 & 0 & 0 & 0 & f_E\\ 0 & -1/R & 0 & 0 & 0 & 0 & 0 & 0 & \omega_N\\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & f_D & 0 & 0 & -f_N\\ 0 & 0 & 0 & 0 & 1/R & 0 & 0 & 0 & -\omega_E\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 2g/R & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}}_{\mathbf{F}(t)} \cdot \underbrace{\begin{bmatrix}\delta N\\ \delta v_N\\ \phi_E\\ \delta E\\ \delta v_E\\ \phi_N\\ \delta D\\ \delta v_D\\ \phi_D\end{bmatrix}}_{\mathbf{x}(t)} + \underbrace{\begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}}_{\mathbf{G}(t)} \cdot \underbrace{\begin{bmatrix}u_{vN}\\ u_{\phi E}\\ u_{vE}\\ u_{\phi N}\\ u_{vD}\\ u_{\phi D}\end{bmatrix}}_{\mathbf{u}(t)}, \tag{12}$$

where: **F** – a fundamental disturbances.

The algorithms of filtration and smoothing presented further on in this paper are discrete, thus they require formulation of a discrete version of the state-space model for a given sampling period $T$. Thus, the above continuous dynamics model must be transformed into its discrete counterpart with the use of known methods presented *e.g.* in [7, 9–11]. The obtained discrete dynamics model is as follows:

$$\underbrace{\begin{bmatrix}\delta N(k+1)\\ \delta v_N(k+1)\\ \phi_E(k+1)\\ \delta E(k+1)\\ \delta v_E(k+1)\\ \phi_N(k+1)\\ \delta D(k+1)\\ \delta v_D(k+1)\\ \phi_D(k+1)\end{bmatrix}}_{\mathbf{x}(k+1)} = \underbrace{\begin{bmatrix}1 & T & -\frac{f_D T^2}{2} & 0 & 0 & 0 & 0 & 0 & \frac{f_E T^2}{2}\\ 0 & 1+\frac{f_D T^2}{2R} & -f_D T & 0 & 0 & 0 & 0 & 0 & f_E T-\frac{f_D \omega_N T^2}{2}\\ 0 & -\frac{T}{R} & 1+\frac{f_D T^2}{2R} & 0 & 0 & 0 & 0 & 0 & -\frac{f_E T^2}{2R}+\omega_N T\\ 0 & 0 & 0 & 1 & T & \frac{f_D T^2}{2} & 0 & 0 & -\frac{f_N T^2}{2}\\ 0 & 0 & 0 & 0 & 1+\frac{f_D T^2}{2R} & f_D T & 0 & 0 & -f_N T-\frac{f_D \omega_E T^2}{2}\\ 0 & 0 & 0 & 0 & \frac{T}{R} & 1+\frac{f_D T^2}{2R} & 0 & 0 & -\frac{f_N T^2}{2R}-\omega_E T\\ 0 & 0 & 0 & 0 & 0 & 0 & 1+\frac{gT^2}{R} & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 2\frac{gT}{R} & 1+\frac{gT^2}{R} & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}}_{\Phi(k+1,k)} \cdot \underbrace{\begin{bmatrix}\delta N(k)\\ \delta v_N(k)\\ \phi_E(k)\\ \delta E(k)\\ \delta v_E(k)\\ \phi_N(k)\\ \delta D(k)\\ \delta v_D(k)\\ \phi_D(k)\end{bmatrix}}_{\mathbf{x}(k)} + \underbrace{\begin{bmatrix}w_N(k)\\ w_{vN}(k)\\ w_{\phi E}(k)\\ w_E(k)\\ w_{vE}(k)\\ w_{\phi N}(k)\\ w_D(k)\\ w_{vD}(k)\\ w_{\phi D}(k)\end{bmatrix}}_{\mathbf{w}(k)}. \tag{13}$$

The observation model of the system describes a relationship between the measurements contained in the vector **z** and the states contained in the vector **x**. In the designed INS/GNSS system the measurements are formed from differences between INS and GNSS position and velocity components, thus they are linearly related to chosen elements of the state vector. The observation model for the described system is given as follows:

$$\underbrace{\begin{bmatrix} R\left[\varphi^{INS}(k)-\varphi^{GNSS}(k)\right] \\ v_N^{INS}(k)-v_N^{GNSS}(k) \\ R\cos\varphi\left[\lambda^{INS}(k)-\lambda^{GNSS}(k)\right] \\ v_E^{INS}(k)-v_E^{GNSS}(k) \\ h^{INS}(k)-h^{GNSS}(k) \\ v_D^{INS}(k)-v_D^{GNSS}(k) \end{bmatrix}}_{z(k)} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{H}(k)} \underbrace{\begin{bmatrix} \delta N(k) \\ \delta v_N(k) \\ \phi_E(k) \\ \delta E(k) \\ \delta v_E(k) \\ \phi_N(k) \\ \delta D(k) \\ \delta v_D(k) \\ \phi_D(k) \end{bmatrix}}_{\boldsymbol{x}(k)} + \underbrace{\begin{bmatrix} v_N(k) \\ v_{vN}(k) \\ v_E(k) \\ v_{vE}(k) \\ v_D(k) \\ v_{vD}(k) \end{bmatrix}}_{\boldsymbol{v}(k)}, \tag{14}$$

where: $\varphi^{INS}$, $\lambda^{INS}$, $h^{INS}$ – INS position coordinates (latitude, longitude, altitude); $\varphi^{GNSS}$, $\lambda^{GNSS}$, $h^{GNSS}$ – GNSS position coordinates; $v_N^{INS}$, $v_E^{INS}$, $v_D^{INS}$ – INS velocity components; $v_N^{GNSS}$, $v_E^{GNSS}$, $v_D^{GNSS}$ – GNSS velocity components; $v_N$, $v_E$, $v_D$, $v_{vN}$, $v_{vE}$, $v_{vD}$ – GNSS measurement errors; $\varphi$ – the true latitude (in practice – approximated by the measured or estimated latitude).

To complete the model of the system it is necessary to calculate the covariance matrix **Q** of the vector **w** of discrete process disturbances and the covariance matrix **R** of the vector **v** of measurement errors. The matrix **Q** has been obtained with the use of the method presented in [9, 11] and is given below with (15) – (34):

$$\boldsymbol{Q} = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & -\frac{f_E f_N T^5}{20}S_{\varphi D} & -\frac{f_E f_N T^4}{8}S_{\varphi D} & Q_{16} & 0 & 0 & \frac{f_E T^3}{6}S_{\phi D} \\ Q_{21} & Q_{22} & Q_{23} & -\frac{f_E f_N T^4}{8}S_{\varphi D} & -\frac{f_E f_N T^3}{3}S_{\varphi D} & Q_{26} & 0 & 0 & \frac{f_E T^2}{2}S_{\phi D} \\ Q_{31} & Q_{32} & Q_{33} & Q_{34} & Q_{35} & Q_{36} & 0 & 0 & Q_{39} \\ -\frac{f_E f_N T^5}{20}S_{\varphi D} & -\frac{f_E f_N T^4}{8}S_{\varphi D} & Q_{43} & Q_{44} & Q_{45} & Q_{46} & 0 & 0 & -\frac{f_N T^3}{6}S_{\phi D} \\ -\frac{f_E f_N T^4}{8}S_{\varphi D} & -\frac{f_E f_N T^3}{3}S_{\varphi D} & Q_{53} & Q_{54} & Q_{55} & Q_{56} & 0 & 0 & -\frac{f_N T^2}{2}S_{\phi D} \\ Q_{61} & Q_{62} & Q_{63} & Q_{64} & Q_{65} & Q_{66} & 0 & 0 & Q_{69} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{T^3 S_{vD}}{3} & \frac{T^2 S_{vD}}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{T^2 S_{vD}}{2} & TS_{vD} & 0 \\ \frac{f_E T^3}{6}S_{\varphi D} & \frac{f_E T^2}{2}S_{\varphi D} & Q_{93} & -\frac{f_N T^3}{6}S_{\varphi D} & -\frac{f_N T^2}{2}S_{\varphi D} & Q_{96} & 0 & 0 & TS_{\phi D} \end{bmatrix}, \tag{15}$$

$$Q_{11} = \frac{T^3}{3}S_{vN} + \frac{f_D^2 T^5}{20}S_{\phi E} + \frac{f_E^2 T^5}{20}S_{\phi D}, \tag{16}$$

$$Q_{12} = Q_{21} = \frac{T^2}{2}S_{vN} + \frac{f_D^2 T^4}{8}S_{\phi E} + \frac{f_E^2 T^4}{8}S_{\phi D}, \tag{17}$$

$$Q_{13} = Q_{31} = -\frac{T^3}{3R}S_{vN} - \frac{f_D T^3}{6}S_{\phi E} + \frac{f_E \omega_N T^4}{8}S_{\phi D}, \tag{18}$$

$$Q_{22} = TS_{vN} + \frac{f_D^2 T^3}{3}S_{\phi E} + \frac{f_E^2 T^3}{3}S_{\phi D}, \tag{19}$$

$$Q_{23} = Q_{32} = -\frac{T^2}{2R}S_{vN} - \frac{f_D T^2}{2}S_{\phi E} + \frac{f_E \omega_N T^3}{3}S_{\phi D}, \tag{20}$$

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 127–142.

$$Q_{33} = \frac{T^3}{3R^2} S_{vN} + TS_{\phi E} + \frac{\omega_N^2 T^3}{3} S_{\phi D}, \tag{21}$$

$$Q_{16} = Q_{61} = -\frac{f_E \omega_E T^4}{8} S_{\phi D} - \frac{f_E f_N T^5}{20R} S_{\phi D}, \tag{22}$$

$$Q_{26} = Q_{62} = -\frac{f_E \omega_E T^3}{3} S_{\phi D} - \frac{f_E f_N T^4}{8R} S_{\phi D}, \tag{23}$$

$$Q_{34} = Q_{43} = -\frac{f_N \omega_N T^4}{8} S_{\phi D} + \frac{f_E f_N T^5}{20R} S_{\phi D}, \tag{24}$$

$$Q_{35} = Q_{53} = -\frac{f_N \omega_N T^3}{3} S_{\phi D} - \frac{f_E f_N T^4}{8R} S_{\phi D}, \tag{25}$$

$$Q_{36} = Q_{63} = -\frac{\omega_E \omega_N T^3}{3} S_{\phi D} + \frac{(f_E \omega_E - f_N \omega_N) T^4}{8R} S_{\phi D} + \frac{f_E f_N T^5}{20R^2} S_{\phi D}, \tag{26}$$

$$Q_{44} = \frac{T^3}{3} S_{vE} + \frac{f_D^2 T^5}{20} S_{\phi N} + \frac{f_N^2 T^5}{20} S_{\phi D}, \tag{27}$$

$$Q_{45} = Q_{54} = \frac{T^2}{2} S_{vE} + \frac{f_D^2 T^4}{8} S_{\phi N} + \frac{f_N^2 T^4}{8} S_{\phi D}, \tag{28}$$

$$Q_{46} = Q_{64} = \frac{T^3}{3R} S_{vE} + \frac{f_D T^3}{6} S_{\phi N} + \frac{f_N \omega_E T^4}{8} S_{\phi D}, \tag{29}$$

$$Q_{55} = TS_{vE} + \frac{f_D^2 T^3}{3} S_{\phi N} + \frac{f_N^2 T^3}{3} S_{\phi D}, \tag{30}$$

$$Q_{56} = Q_{65} = \frac{T^2}{2R} S_{vE} + \frac{f_D T^2}{2} S_{\phi N} + \frac{f_N \omega_E T^3}{3} S_{\phi D}, \tag{31}$$

$$Q_{66} = \frac{T^3}{3R^2} S_{vE} + TS_{\phi N} + \frac{\omega_E^2 T^3}{3} S_{\phi D}, \tag{32}$$

$$Q_{39} = Q_{93} = \frac{\omega_N T^2}{2} S_{\phi D} - \frac{f_E T^3}{6R} S_{\phi D}, \tag{33}$$

$$Q_{69} = Q_{96} = -\frac{\omega_E T^2}{2} S_{\phi D} - \frac{f_N T^3}{6R} S_{\phi D}, \tag{34}$$

where: $S_{vN}, S_{\phi E}, S_{vE}, S_{\phi N}, S_{vD}, S_{\phi D}$ – power spectral densities of Gaussian white noise in the vector **u** of continuous random process disturbances.

The measurement errors of GNSS receiver have been for simplicity modelled as uncorrelated in time and between each other Gaussian random sequences of zero mean and constant variance. As a result, the covariance matrix of measurement errors **R** is diagonal and is given as follows:

$$\mathbf{R} = \begin{bmatrix} \sigma_N^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{vN}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_E^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{vE}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_D^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{vD}^2 \end{bmatrix}, \tag{35}$$

where: $\sigma_N^2$, $\sigma_E^2$, $\sigma_D^2$ – variances of position errors of the GNSS receiver; $\sigma_{vN}^2$, $\sigma_{vE}^2$, $\sigma_{vD}^2$ – variances of velocity errors of the GNSS receiver.

## 3. Kalman filtering algorithm

In applications requiring on-line estimation of the state-vector $\mathbf{x}(k)$ without delays, various filtering algorithms of the incoming navigation measurements are usually applied. The problem of filtering consists in finding state estimates of $\mathbf{x}(k)$ for all time steps $k$ on the basis of all measurements made up to this time. Such an estimate is given as follows:

$$\hat{\mathbf{x}}(k \mid k) = E[\mathbf{x}(k) \mid \mathbf{z}(1), \ldots, \mathbf{z}(k)]. \tag{36}$$

For linear systems, the optimal filtering algorithm is the linear Kalman filter [5, 9, 19]. Due to the linearity of the formulated model of the INS/GNSS integrated navigation system, the linear Kalman filter has been chosen as one of the algorithms to be designed and implemented in the system. A block diagram of the algorithm is presented in Fig. 1. It contains the initialization (step 1), executed once at the beginning of filter's operation, and recursively executed steps of time update (step 2), acquiring a new measurement (step 3) and a measurement update (step 4).

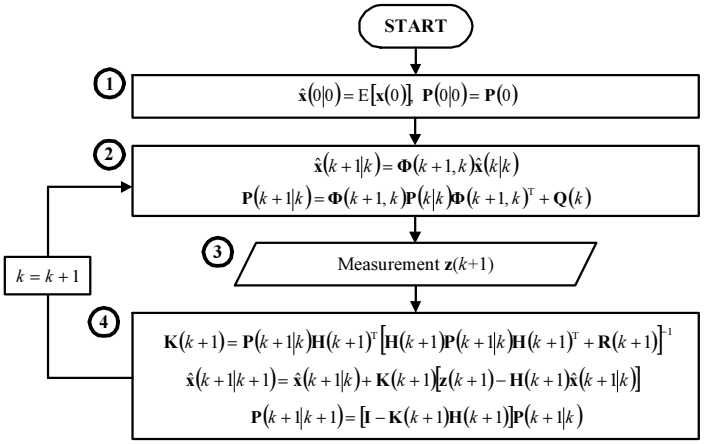



Fig. 1. The Kalman filtering algorithm.

The symbols used in the above diagram are as follows: $\hat{\mathbf{x}}(k+1 \mid k)$ – a predicted state vector in the time step $k + 1$; $\hat{\mathbf{x}}(k+1 \mid k+1)$ – corrected state vector in the time step $k + 1$; $\mathbf{P}(k + 1|k)$ – a covariance matrix of prediction errors; $\mathbf{P}(k + 1|k + 1)$ – a covariance matrix of filtration errors; $\mathbf{K}(k + 1)$ – a Kalman gains' matrix; $\mathbf{I}$ – an identity matrix. The other matrices used in the equations come from the previously defined state-space model of the system.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 127–142.

## 4. Smoothing algorithms

Apart from the above Kalman filter, two smoothing algorithms have been developed, *i.e.* a fixed-interval algorithm for off-line estimation and a fixed-lag algorithm for on-line estimation of position, velocity and orientation of UAV. The operation of smoothing consists in estimation of the state-space vector $\mathbf{x}(k)$ in the time step $k$ on the basis of measurements from time steps later than $k$. Thus, it can be accomplished after mission of UAV or during its flight but with a short delay.

### *4.1. Fixed-interval smoothing algorithm*

In the fixed-interval smoothing we assume that the measurements gathered in an interval $[0, N]$ are known. In our system they are registered on board UAV during flight. The algorithm is responsible for finding optimal state estimates of $\mathbf{x}(k)$ for all time steps $k$ inside this interval on the basis of all known measurements. Such an estimate is given as follows:

$$\hat{\mathbf{x}}(k \mid N) = E[\mathbf{x}(k) \mid \mathbf{z}(1), \ldots, \mathbf{z}(N)], \tag{37}$$

for $k = 0,1, \ldots, N$. As the estimate is based on all available measurements, a properly executed fixed-interval smoothing provides the best possible estimate of the state vector.

There exist several methods of fixed-interval smoothing. One of the most commonly applied is an algorithm proposed by Rauch, Tung and Striebel [10, 19, 21, 22], known as the RTS algorithm. It is accomplished in two consecutive stages, *i.e.* forward and backward filtering. The forward filtering consists in calculation of estimates of the state vector $\mathbf{x}(k)$ with the use of the optimal KF. The results obtained in each time step $k$ have to be registered for further use. It is necessary to store the estimates of the state vector obtained during filtration $\hat{\mathbf{x}}(k \mid k)$ and one-step prediction $\hat{\mathbf{x}}(k+1 \mid k)$ as well as their error covariance matrices $\mathbf{P}(k \mid k)$ and $\mathbf{P}(k+1 \mid k)$. In non-stationary systems, also variable values of the transition matrix $\mathbf{\Phi}(k+1,k)$ have to be stored. After this first stage of data processing, the backward filtering is accomplished with the initial conditions $\hat{\mathbf{x}}(N \mid N)$ and $\mathbf{P}(N \mid N)$, obtained as the final results of the forward filtering.

The optimal estimate of the state vector $\mathbf{x}(k)$ obtained during the fixed-interval smoothing is given as follows:

$$\hat{\mathbf{x}}(k \mid N) = \hat{\mathbf{x}}(k \mid k) + \mathbf{A}(k)[\hat{\mathbf{x}}(k+1 \mid N) - \hat{\mathbf{x}}(k+1 \mid k)], \tag{38}$$

where $\mathbf{A}(k)$ is the smoothing gain matrix:

$$\mathbf{A}(k) = \mathbf{P}(k \mid k)\mathbf{\Phi}^{\mathrm{T}}(k+1,k)\mathbf{P}^{-1}(k+1 \mid k) \quad \text{for } k = N-1, N-2, \ldots, 0. \tag{39}$$

The error covariance matrix of the fixed-interval smoothing is as follows:

$$\mathbf{P}(k \mid N) = \mathbf{P}(k \mid k) + \mathbf{A}(k)[\mathbf{P}(k+1 \mid N) - \mathbf{P}(k+1 \mid k)]\mathbf{A}^{\mathrm{T}}(k), \tag{40}$$

for $k = N-1, N-2, \ldots, 0$. The idea of fixed-interval smoothing is explained in Fig. 2.

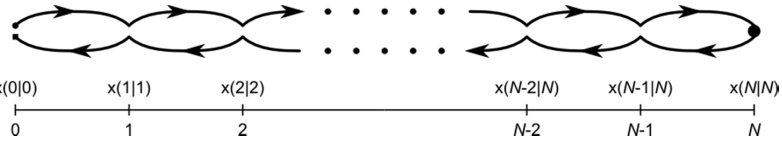



Fig. 2. The idea of fixed-interval smoothing.

The RTS algorithm is easy to implement but its drawback is a necessity of time-consuming inversions of the covariance matrix of prediction errors in (39). Other, less time-consuming fixed-interval smoothing algorithms can be found in literature [19, 23].

### *4.2. Fixed-lag smoothing algorithm*

The fixed-lag smoothing algorithm processes incoming measurements on-line and calculates estimates of the state vector $\mathbf{x}(k)$ for time steps $k$ delayed by a constant number of $N$ steps in comparison with the current measurement. Such an estimate is given as follows:

$$\hat{\mathbf{x}}(k \mid k+N) = E[\mathbf{x}(k) \mid \mathbf{z}(1), \ldots, \mathbf{z}(k), \mathbf{z}(k+1), \ldots, \mathbf{z}(k+N)], \tag{41}$$

for $k = 0,1,2, \ldots$ . The results of fixed-lag smoothing are less accurate than those of fixed-interval smoothing, since its estimates are based on a smaller amount of data. However, for large values of $N$, the accuracy of fixed-lag smoothing approaches the accuracy of the fixed-interval one, which will be demonstrated further on. Moreover, a possibility of using this algorithm on-line, during the flight of UAV, may be an important advantage in many applications.

The optimal estimates of the state vector in the fixed-lag smoothing are formed with the use of the following equation [10, 19]:

$$\hat{\mathbf{x}}(k+1-i \mid k+1) = \hat{\mathbf{x}}(k+1-i \mid k) + \mathbf{K}_i(k+1)\tilde{\mathbf{z}}(k+1), \tag{42}$$

for $i = 1,2, \ldots, N$, where $\mathbf{K}_i(k+1)$ represents the gain matrix of the optimal fixed-lag smoother and it can be calculated as presented in [19]. The smoothing algorithm uses estimates of the state vector and residuals $\tilde{\mathbf{z}}(k+1)$ from a Kalman filter designed for the original state-space model. Thus, such a Kalman filter must be implemented to provide the data for the fixed-lag smoother. The smoothing algorithm can be accomplished in parallel to the Kalman filter.

As the estimate from the fixed-lag smoothing algorithm is delayed by $N$ time steps, we can consider the smoothing process as accomplished in a time window of length $N$. This window is moving forward along the time scale as new measurements are processed. The idea of fixed-lag smoothing in a moving time window is explained in Fig. 3.

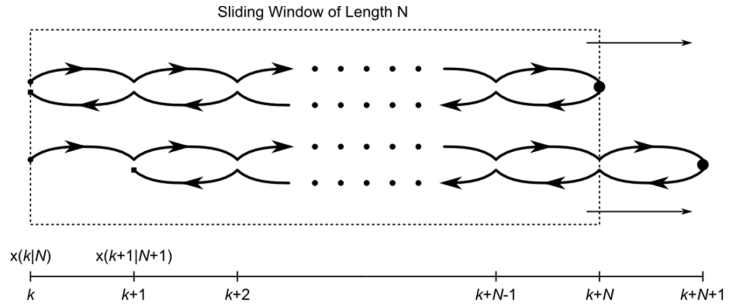



Fig. 3. The idea of fixed-lag smoothing.

The fixed-lag smoothing algorithm used in this paper is more complicated than the fixed-interval one as it requires numerous matrix multiplications in the process of calculation of the gain matrix of the optimal fixed-lag smoother $\mathbf{K}_i(k+1)$. If necessary, a simpler solution, where a fixed-interval smoothing is used to solve the fixed-lag smoothing problem can be found in the literature [18, 21].

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 127–142.

## 5. Simulation results

The Kalman filter as well as both presented smoothing algorithms have been implemented in the assumed model of INS/GNSS system and simulated with the use of Matlab®. A Matlab toolbox called IRENA, developed at the Institute of Radioelectronics, at the Military University of Technology, Warsaw, Poland, have been used for this purpose [24]. The toolbox extends the Matlab functionality with useful functions enabling to simulate integrated navigation systems and their components.

During the simulations, a trajectory of flight of UAV, lasting 400 seconds, has been generated and used as a reference in testing filtering and smoothing algorithms. Then, INS and GNSS errors have been generated and added to the reference positions and velocities of UAV. For simplicity, the influence of internal Kalman filter, which is typically implemented in GNSS receivers, has been neglected and GNSS errors have been assumed to be Gaussian zero-mean, constant-variance white noise. Such an omission affects both filtering and smoothing, thus the comparisons of both types of algorithms do not affect their validity [25].

The parameters of INS and GNSS errors in the simulations have been chosen on the basis of technical specifications of real navigation devices used in the integrated system developed during the WATSAR project. These devices include an inertial 1750 IMU measurement unit from KVH Industries and a GNSS receiver built into an INS/GNSS(RTK) Ekinox-D system from SBG Systems. The values of the assumed parameters are given in Table 1.

Table 1. The values of parameters of INS and GNSS errors assumed in simulations.

| Parameter | Value |
|---|---|
| $S_{\phi N}$, $S_{\phi E}$, $S_{\phi D}$ | $1.15 \cdot 10^{-11}$ rad²/s |
| $S_{vN}$, $S_{vE}$, $S_{vD}$ | $1.4 \cdot 10^{-6}$ m²/s³ |
| $\sigma_N$, $\sigma_E$, $\sigma_D$ | 1.2 m (SP), 0.4 m (DGNSS) |
| $\sigma_{vN}$, $\sigma_{vE}$, $\sigma_{vD}$ | 0.02 m/s |

The parameters $S_{\phi N}$, $S_{\phi E}$, $S_{\phi D}$ represent power spectral densities of errors of gyros, whereas $S_{vN}$, $S_{vE}$, $S_{vD}$ are power spectral densities of errors of accelerometers composing INS. The parameters $\sigma_N$, $\sigma_E$, $\sigma_D$ represent standard deviations of GNSS position errors, and $\sigma_{vN}$, $\sigma_{vE}$, $\sigma_{vD}$ are standard deviations of GNSS velocity errors expressed in the NED system of coordinates. The simulations have been performed for two different accuracy levels of GNSS possible in our system: the *standard positioning* (SP) accuracy and the accuracy of GNSS with differential corrections (DGNSS) [5]. A period of availability of new GNSS data has been assumed to be 0.5 second. The parameters given in Table 1 have also been used in implementation of the Kalman filter.

In the first step of simulations, the positioning errors of INS/GNSS system for SP and DGNSS levels of GNSS accuracy, with the Kalman filter and with the fixed-interval smoothing algorithm have been compared. These errors are expressed in the local horizontal NED reference system and their chosen results for the DGNSS level of accuracy are presented in Fig. 4–6.

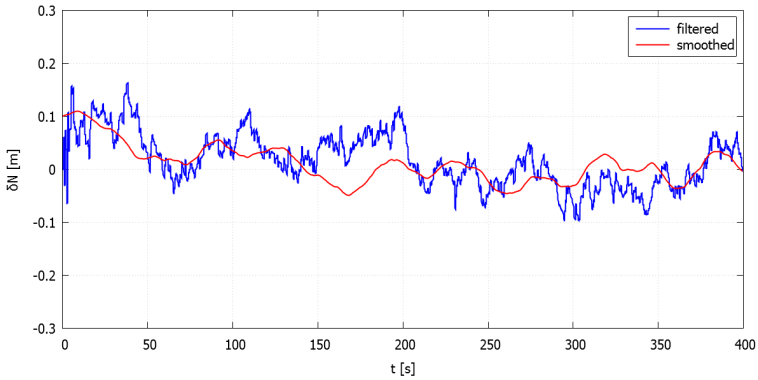



Fig. 4. The positioning errors in INS/GNSS system in the north direction.

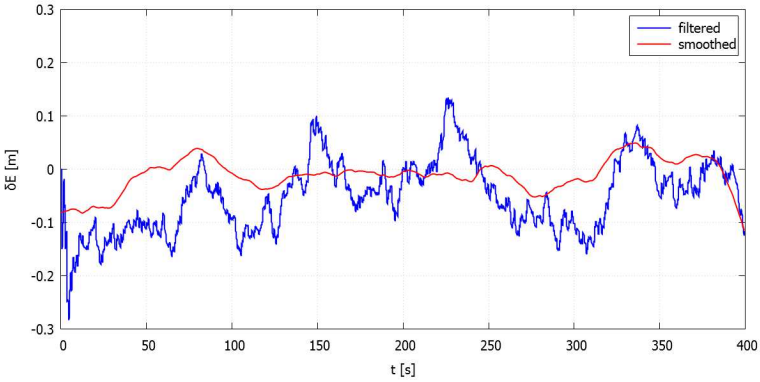



Fig. 5. The positioning errors in INS/GNSS system in the east direction.

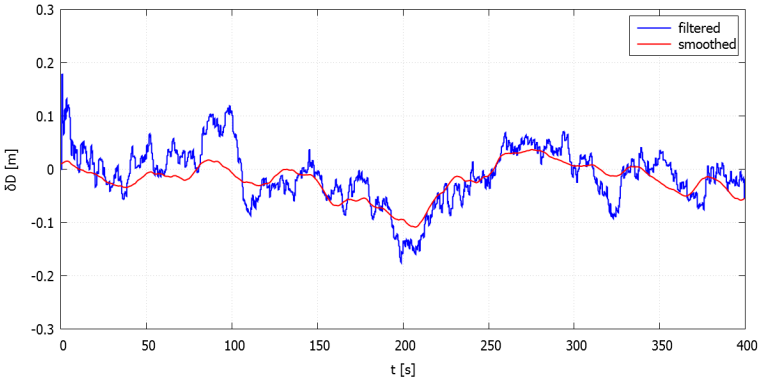



Fig. 6. The positioning errors in INS/GNSS system in the vertical (down) direction.

The above results show that the smoothed estimates of position are more accurate than the estimates from the Kalman filter. The level of improvement can be quantitatively assessed when we compare theoretical errors for various states estimated by both algorithms. The Kalman filter provides such information as its standard equations include a calculation of the error covariance matrix of filtration $\mathbf{P}(k \mid k)$ in each time step $k$. The smoothing algorithms do not include

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 127–142.

or require such calculations, however, the error covariance matrix of smoothing can also be easily evaluated [9, 10, 19], which has been done for the purpose of comparisons. The diagonal elements of the error covariance matrices represent theoretical variances of estimation errors of respective states and their square roots are standard deviations of these errors. The comparison of theoretical standard deviations of positioning errors in the north direction for the entire period of simulations is shown in Fig. 7. Similar results have been obtained for other components of the state vector, therefore they are not included in the paper. From Fig. 7 we can see that, apart from the initial and final intervals of simulations, lasting 20 seconds each, the fixed-interval smoothing is about twice more accurate than the Kalman filter.

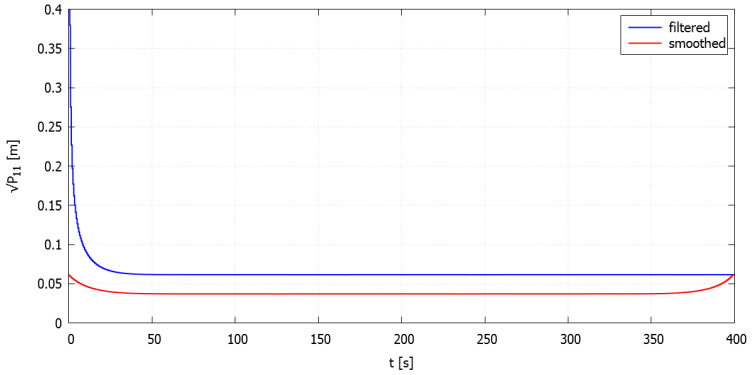



Fig. 7. The theoretical standard deviations of positioning errors in the north direction.

The comparison of filtering and fixed-interval smoothing accuracy can also be performed when calculating *root mean-squared* (RMS) errors of positioning for the whole period of simulation. The RMS errors of position for all axes of the NED reference system, as well as the total RMS positioning error calculated according to the following formula:

$$RMS(\delta_P) = \sqrt{[RMS(\delta_N)]^2 + [RMS(\delta_E)]^2 + [RMS(\delta_D)]^2} \tag{43}$$

for SP and DGNSS levels of GNSS accuracy are presented in Table 2.

Table 2. The RMS errors of positioning in INS/GNSS system with a Kalman filter and a fixed-interval smoother.

| Errors [m] | SP | | | DGNSS | | |
|---|---|---|---|---|---|---|
| | Filtering | Smoothing | Reduction | Filtering | Smoothing | Reduction |
| RMS($\delta_N$) | 0.131 | 0.072 | 45.6% | 0.067 | 0.038 | 43.6% |
| RMS($\delta_E$) | 0.134 | 0.073 | 45.4% | 0.069 | 0.037 | 45.4% |
| RMS($\delta_D$) | 0.127 | 0.068 | 46.3% | 0.065 | 0.036 | 45.0% |
| RMS($\delta_P$) | 0.226 | 0.123 | 45.8% | 0.115 | 0.064 | 44.7% |

The above results prove that the fixed-interval smoothing significantly reduces errors of position estimation in comparison with the Kalman filtering. Thus, for the assumed parameters of navigation devices, the off-line reconstruction of the UAV trajectory with the use of a fixed-interval smoother can be about twice as accurate as that with the use of a Kalman filter. This result is in accordance with the previously presented comparison of theoretical standard deviations of positioning errors (Fig. 7). The effects of error reduction are similar for all the coordinates and for both levels of GNSS accuracy.

In the next step of simulations, the positioning errors of INS/GNSS system for SP and DGNSS levels of GNSS accuracy, with a Kalman filter, fixed-interval and fixed-lag smoothing algorithms for various lags $N$ have been compared. As a period between time steps in simulations is equal to 0.5 second, the time delay of the smoothed estimate is equal to $\Delta t = N/2$ seconds. The behavior of positioning errors follows a similar pattern for all the axes, therefore only the errors along the north axis, for the DGNSS level of accuracy, have been chosen for presentation and shown in Fig. 8–13.

The total RMS positioning errors for SP and DGNSS levels of GNSS accuracy for the Kalman filter, the fixed-interval smoother and the fixed-lag smoother with various delays are presented in Table 3. The error reduction in comparison with the filtering is shown in brackets. It is worth to notice that the relative level of error reduction asymptotically approaches a level equal to that of the fixed-interval smoother for both SP and DGNSS. However, when the correcting device (GNSS receiver) is more accurate, the progress of this reduction is quicker. Thus, the fixed-lag smoothing requires less delay to approach the quality of the fixed-interval one when a more accurate correcting sensor is used in the integrated navigation system.

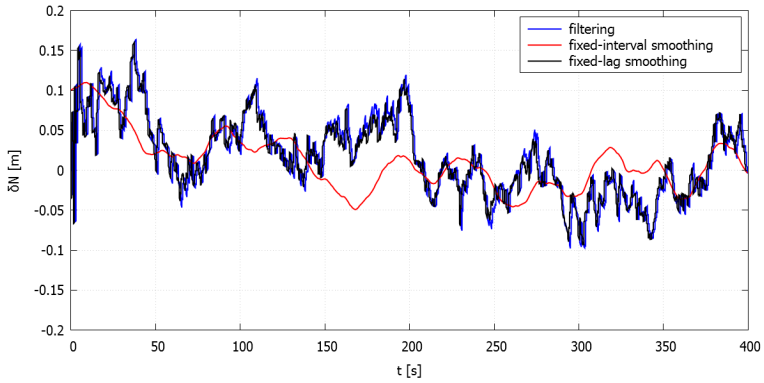



Fig. 8. The positioning errors in INS/GNSS system in the north direction ($N = 1$, $\Delta t = 0.5$ s, in the fixed-lag smoothing).

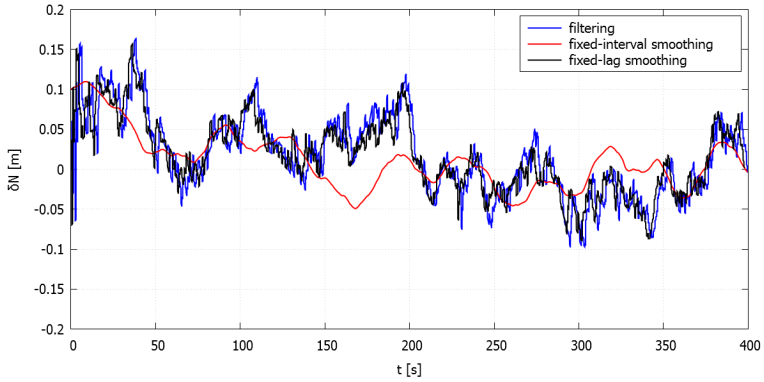



Fig. 9. The positioning errors in INS/GNSS system in the north direction ($N = 2$, $\Delta t = 1$ s, in the fixed-lag smoothing).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 127–142.

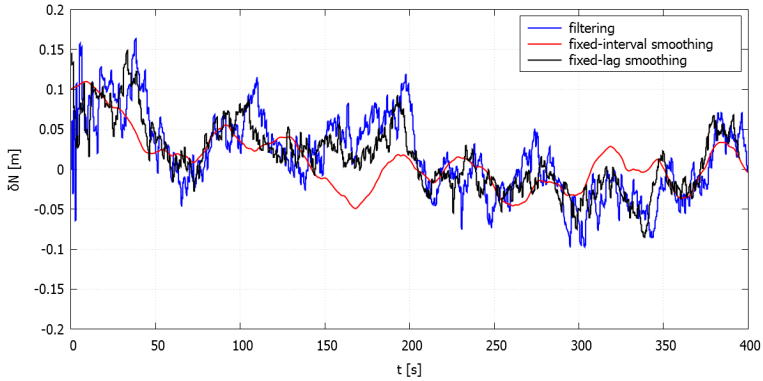



Fig. 10. The positioning errors in INS/GNSS system in the north direction ($N = 5$, $\Delta t = 2.5$ s, in the fixed-lag smoothing).

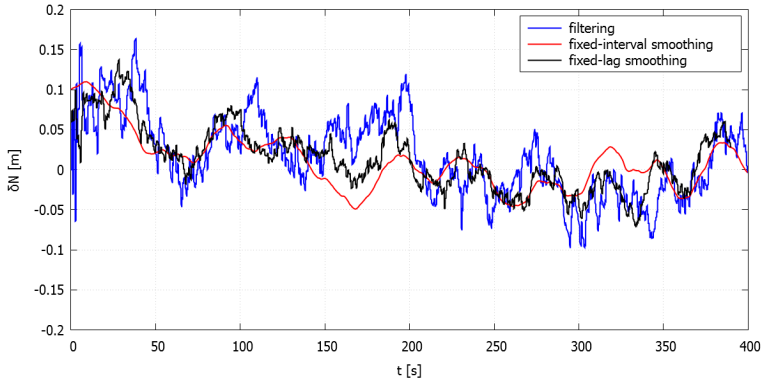



Fig. 11. The positioning errors in INS/GNSS system in the north direction $N = 10$, $\Delta t = 5$ s, in the fixed-lag smoothing).

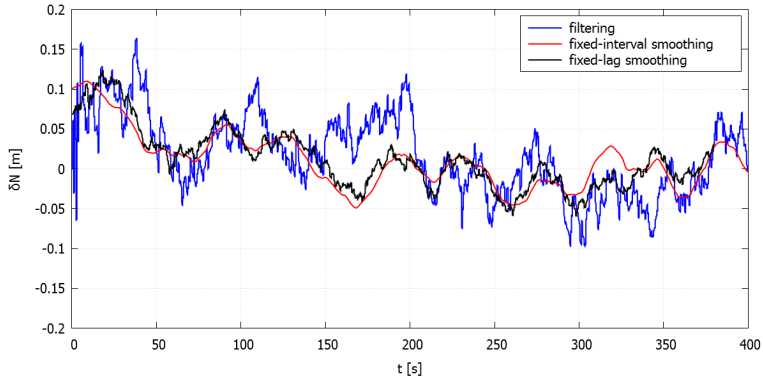



Fig. 12. The positioning errors in INS/GNSS system in the north direction ($N = 20$, $\Delta t = 10$ s, in the fixed-lag smoothing).

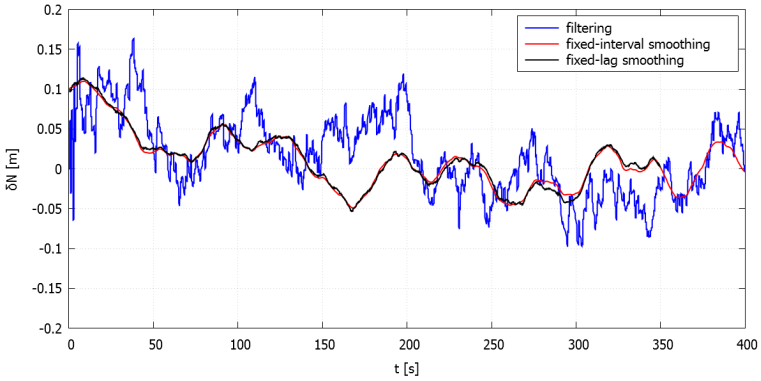



Fig. 13. The positioning errors in INS/GNSS system in the north direction ($N = 50$, $\Delta t = 25$ s, in the fixed-lag smoothing).

Table 3. The RMS errors of positioning in INS/GNSS system with a Kalman filter and a fixed-lag smoother.

| RMS($\delta_P$) [m] | Filtering | Fixed-lag smoothing for various delays | | | | | | Fixed-interval smoothing |
|---|---|---|---|---|---|---|---|---|
| | | $N = 1$ | $N = 2$ | $N = 5$ | $N = 10$ | $N = 20$ | $N = 50$ | |
| | | $\Delta t = 0.5$s | $\Delta t = 1$s | $\Delta t = 2.5$s | $\Delta t = 5$s | $\Delta t = 10$s | $\Delta t = 25$s | |
| **SP** | 0.226 | 0.217 (2.3%) | 0.211 (5.4%) | 0.197 (11.7%) | 0.183 (18.2%) | 0.164 (26.7%) | 0.138 (39%) | 0.123 (45.8%) |
| **DGNSS** | 0.115 | 0.109 (4.8%) | 0.104 (9.2%) | 0.093 (18.2%) | 0.083 (27.6%) | 0.072 (37.5%) | 0.064 (44.4%) | 0.064 (44.7%) |

## 6. Conclusions

The results presented in this paper demonstrate that both fixed-interval and fixed-lag smoothing algorithms can be very useful in specific navigation applications. A fixed-interval smoother can be used in post-processing of registered navigation data, *e.g.* for off-line reconstruction of the trajectory and parameters of flight of a UAV. In such an application, the accuracy of smoother is significantly better than the accuracy of a Kalman filter, which is typically used for this purpose. For the assumed parameters of devices, the errors of fixed-interval smoothing have been about twice smaller than the errors of filtering.

On the other hand, a fixed-lag smoother can be used instead of a Kalman filter for on-line estimation of position, velocity and orientation of a UAV, in applications accepting relatively small delay of the output data. Such applications include *e.g.* synthetic aperture radars which are an important type of image intelligence systems of today. The results presented in this paper demonstrate that a fixed-lag smoothing algorithm is more accurate than a Kalman filter. Its accuracy increases along with the increasing delay of estimates. Moreover, the accuracy of a fixed-lag smoother asymptotically approaches that of a fixed-interval one and makes it in a relatively short time. In the case of our system, it requires only several tens of seconds of delay, which can be acceptable in many applications.

It is important to notice that the use of a more accurate correcting device or a more accurate mode of its operation (*e.g.* DGNSS instead of SP in the case of a GNSS receiver) shortens the time necessary to achieve the required level of reduction of errors and a fixed-lag smoother can achieve the same level of accuracy with shorter delays.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 127–142.

## Acknowledgements

This project was supported by the National Centre for Research and Development, Poland, within the scope of Applied Research Programme under Research Project PBS/B3/15/2012.

## References

[1] Matuszewski, J. (2008). Specific emitter identification. *International Radar Symposium*, Wroclaw, 1–4.

[2] Cumming, I.G., Wong, F.H., (2005). *Digital Processing of Synthetic Aperture Radar Data. Algorithms and Implementation*. Artech House.

[3] Mengdao, X., Xiuwei, J., Renbiao, W., Feng, Z., Zheng, B. (2009). Motion Compensation for UAV SAR Based on Raw Radar Data. *IEEE Transactions an Geoscience and Remote Sensing*, 8, 2870–2883.

[4] Samczyński, P., Malanowski, M., Gromek, D., Gromek, A., Kulpa, K., Krzonkalla, J., Mordzonek, M., Nowakowski, M. (2014). Effective SAR image creation using low cost INS/GPS. *International Radar Symposium*, Gdańsk, 174–177.

[5] Groves, P. (2008). *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems*. Artech House.

[6] Titterton, D.H., Weston, J.L. (2004). Strapdown Inertial Navigation Technology. *Institution of Electrical Engineers*, UK, 17–57.

[7] Farrell, J.A. (2008). *Aided Navigation GPS with High Rate Sensors*. McGraw-Hill.

[8] Anderson, B.D.O., Moore, J.B. (1979). *Optimal Filtering*. Prentice-Hall, INC, Englewood Cliffs, New Jersey.

[9] Brown, R.G., Hwang, P.Y.C. (2012). *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley & Sons, Inc.

[10] Särkkä, S. (2013). *Bayesian filtering and smoothing*. Cambridge University Press.

[11] Kaniewski, P. (2010). *Structures, models and algorithms in integrated positioning and navigation systems*. Wyd. WAT, Warszawa.

[12] Gelb, A. (2001). *Applied Optimal Estimation*. Massachusetts Institute of Technology, Cambridge, Massachusetts and London.

[13] Li, X., Xie, Y., Bi, D., Ao, Y. (2013). Kalman Filter Based Method for Fault Diagnostics of Analog Circuits. *Metrol. Meas. Syst.*, 20(2), 307–322.

[14] Konatowski, S., Pieniężny, A.T. (2007). A comparison of estimation accuracy by the use of KF, EKF & UKF filters. *WIT Transactions on Modelling and Simulation*, 46, 779–789.

[15] Śmieszek, M., Dobrzańska, M. (2015). Application of Kalman Filter in Navigation Process of Automated Guided Vehicles. *Metrol. Meas. Syst.*, 22(3), 443–454.

[16] Fornaro, G. (1999). Trajectory deviations in airborne SAR: analysis and compensation. *IEEE Transactions on Aerospace and Electronics Systems*, 35(3), 997–1009.

[17] Kaniewski, P., Leśnik, C., Susek, W., Serafin, P. (2015). Airborne radar terrain imaging system. *International Radar Symposium*, Dresden, 248–253.

[18] Einicke, G.A. (2012). *Smoothing, Filtering and Prediction: Estimating the Past, Present and Future*. Published by InTech.

[19] Meditch, J.S. (1969). *Stochastic Optimal Linear Estimation and Control*. New York, McGraw Hill.

[20] Łabowski, M., Kaniewski, P., Konatowski, S., (2016). Estimation of flight path deviations for SAR radar installed on UAV. *Metrol. Meas. Syst.*, 23(3), 383–391.

[21] Rauch, H.E., (1963). Solutions to the Linear Smoothing Problem. *IEEE Transactions on Automatic Control*, 8, 371–372.

[22] Rauch, H.E., Tung, F., Striebel, C.T. (1965). Maximum Likelihood Estimation of Linear Dynamic Systems. *AIAA Journal*, 3(8), 1445–1450.

[23] Techy, L., Morgansen, K.A., Woolsey, C.A. (2011). Long-baseline acoustic localization of the Seaglider underwater glider. *American Control Conference (ACC)*, San Francisco.

[24] Kaniewski, P., Konatowski, S., (2014). Software Toolbox for Simulation of Integrated Navigation Systems. *Przegląd Elektrotechniczny*, 90(8), 168–171.

[25] Bednarek, M., Będkowski, L. (2008). Dąbrowski T.: Comparative-threshold diagnosing in messages transmission system. *Przegląd Elektrotechniczny*, 84(11A), 320–324.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 143–151.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# EFFECTS OF RADIATION DOSES ON THE PHOTOSTIMULATED LUMINESCENCE RESPONSE OF CERTAIN HERBS AND SPICES

**Ivana Sandeva, Hristina Spasevska, Margarita Ginovska, Lihnida Stojanovska-Georgievska**

*Ss. Cyril and Methodius University, Faculty of Electrical Engineering and Information Technologies, Ruger Boskovic b.b., 574, 1000 Skopje, Republic of Macedonia (✉ ivana@feit.ukim.edu.mk, +38 923 099 178, hristina@feit.ukim.edu.mk, gmarga@feit.ukim.edu.mk, lihnida@feit.ukim.edu.mk)*

**Abstract**

Ionizing radiation applied on food eliminates harmful microorganisms, prevents sprouting and delays ripening. All methods for detection of irradiated food are based on physical, chemical, biological or microbiological changes caused by the treatment with ionizing radiation. When minerals are exposed to ionizing radiation, they accumulate radiation energy and store it in the crystal lattice, by which some electrons remain trapped in the lattice. When these minerals are exposed to optical stimulation, trapped electrons are released. The phenomenon, called optically stimulated luminescence or photostimulated luminescence, occurs when released electrons recombine with holes from luminescence centers in the lattice, resulting in emission of light with certain wavelengths.
In this paper, the results of measurements performed on seven different samples of herbs and spices are presented. In order to make a comparison between luminescence signals from samples treated with different doses, unirradiated samples are treated with Co-60 with doses of 1 kGy, 5 kGy and 10 kGy. In all cases it was shown that the higher the applied dose, the higher the luminescence signal.

Keywords: food, ionizing radiation, photostimulated luminescence.



## 1. Introduction

Different technologies and methods have been developed and applied to enhance food quality. One of them is the method in which food is exposed to carefully controlled doses of ionizing radiation. This technology destroys harmful microorganisms, prevents sprouting and delays ripening in some fruits and vegetables [1]. Irradiation reduces the use of pesticides and preservatives. This process involves almost no heat, so irradiated foodstuffs remain raw after the treatment. Ionizing radiation is the reason for formation of radicals in food. These radicals can have a direct effect on microorganisms in food with destabilization of DNA, or an indirect effect by forming ions of water which destroy cell components [2]. Different kinds of food are treated with different doses of radiation, which depends on the desired results and type of food. Studies have shown that doses below 10 kGy are not susceptible to toxicological hazard [3]. In some cases plastic packaging may lead to migration of harmful components into foodstuffs [4], so the kind of material used for packaging has to be taken into account.

Development of this method for food preservation, commercial use of ionizing radiation for enhancing food quality and trade with treated food arise the need of reliable and routine tests for detection of irradiated food.

All methods for detection of irradiated food may be classified in three groups: biological, chemical and physical methods. The ideal method for detection should measure a specific effect of the ionizing radiation, which is proportional to the dose and should not depend on the processing and storage of food and the length of time between the treatment and the testing [5].

Article history: received on Mar. 03, 2016; accepted on Aug. 28, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0003.

Physical methods, including measuring impedance, viscosity, thermal analysis, nuclear magnetic resonance, electron spin resonance, luminescence, are based on detection of changes in physical properties [6]. One simple physical method for detection of irradiated food is optically stimulated luminescence or *photostimulated luminescence* (PSL). Samples are stimulated with a pulsed infra-red radiation, and the signal is detected at the end of each pulse. This method gives reliable results only if the food contains sufficient amount of minerals [7–9]. Minerals are substances that are responsible for storage of energy in the defects of their crystal lattice as a result of exposure to ionizing radiation. Optical stimulation releases electrons that have been trapped in the lattice. These electrons return from the excited state to the ground state by losing part of their excess energy as photons, thus a signal could be observed and measured as a luminescence response.

All standardized methods for detection of irradiated food are qualitative, which means that the absorbed doses cannot be precisely determined after treatment. The shape and quantity of the foodstuffs during the treatment, as well as the type and quantity of minerals in food, can affect the result. Efforts are being made to estimate the doses of ionising radiation by which certain foodstuffs have been treated [10–11]. In order to establish fast and qualitative estimation of the applied dose during the treatment we studied the influence of dose on the luminescence signal. Studied samples were treated with doses of 1 kGy, 5 kGy and 10 kGy.

Analytical detection of irradiation processing of food is very important to implement Quality Control of treated food at all levels. Currently, national legislation is based on respecting European Directives, their harmonization with national legislation and other laws and rules. Detection of irradiated food in Europe is regulated under European Legislation L66/16-25 (1999), covering Directives 1999/2/EC and 1999/3/EC [12–13]. Regulation for specific safety requirements of the food treated with ionizing radiation (Official Gazette of Republic of Macedonia, No. 63/2014) has been adopted on the basis of Article 8 paragraph 1 of the Law of food safety and products and materials that come in contact with food, of the Statute of Republic of Macedonia (Official Gazette of Republic of Macedonia, No. 54/2002 and 84/2007) [14–15].

## 2. Experimental

### *2.1. Description of equipment*

Detection of irradiated food is performed by the SUERC *pulsed photostimulated luminescence* (PPSL) system designed and developed at the Scottish Universities Research and Reactor Centre. The procedure has been set according to EU standard for detection of irradiated food – EN 13751:2002, Foodstuffs – Detection of irradiated food using Photostimulated Luminescence. Equipment consists of two main parts: a control unit and a detector head. The detection technique of the system is shown in Fig. 1 and Fig. 2.

Figure 2 illustrates the relative timing of the stimulation light source and photon counter during a preset 15 second test period when the system is operating in the screening mode.

When the IR LEDs are on, the photon counter accumulates PSL signal counts from the test sample, plus the system background counts which are principally due to dark counts from the PMT. While the IR LEDs are off, the system background counts are subtracted from the accumulated counts. This ‘Up-Down’ count system minimizes the effect of the system background signal, thereby increasing the dynamic range and system detectivity for weak PSL emitters [16].

The PMT dark count is temperature-sensitive and approximately doubles for every 5ºC rise. Fluctuation in the dark count is described by a Poisson distribution; hence the statistical

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 143–151.

variation in the dark count rate is proportional to the square root of its mean. Optimum system performance is obtained when the equipment is kept at ambient temperatures.

When testing with an empty sample chamber, the accumulated count may fall below zero. The reason for this are the variations in Up and Down counts due to either the statistical nature of the background count or to small changes in the relative durations of the Up and Down periods. To ensure that this does not happen, the photon counter is pre-loaded with 256 counts at the start of each measurement.

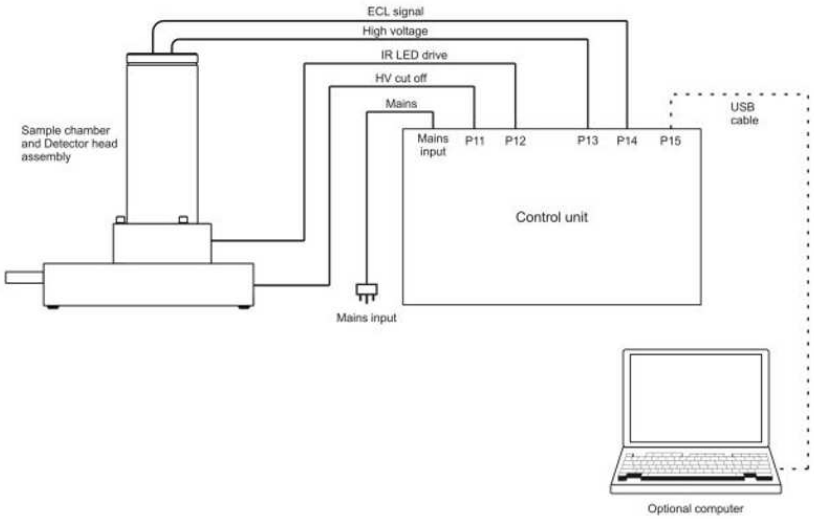



Fig. 1. The Interconnection Diagram of the SUERC PPSL System [16].

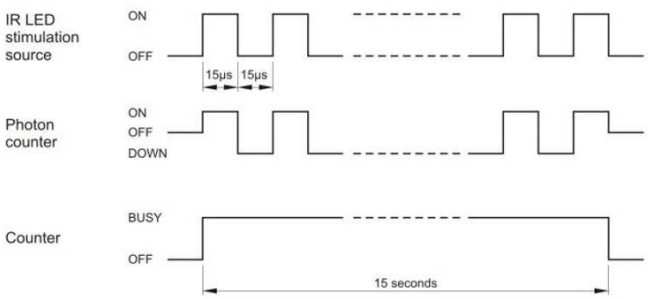



Fig. 2. The time distribution and shape of the signal of the SUERC PPSL System [16].

The main application of this system is rapid screening/detection of irradiated herbs, spices and seasonings, and may be used either in the stand-alone mode or in conjunction with a computer for data storage [16]. The wavelength of stimulating light is in the interval of 450–950 nm, and the obtained signal has wavelengths in the interval of 300–350 nm. The whole system is portable and intended for indoor use. The method is cheap, rapid and applicable for a wide range of foodstuffs. Samples can be tested with almost no preparation and without any damage to them. During measurements samples are stimulated by an array of infra-red light emitting diodes which are pulsed symetrically on and off for equal periods. Luminescence is measured using a patented digital lock-in photon counting method by using a cathode photomultiplier tube (type: 9814B02 from the manufacturer ET Enterprises Limited). This is

a bi-alkali, high-gain, low-background photomultiplier tube. Optical filtering is used to define both the stimulation and detection wavebands. Irradiated samples produce a specific signal which is detected and quantified. The registered signal level is compared with two reference threshold values [17]. Most irradiated samples produce signals above the high threshold, and most unirradiated samples produce signals below the low threshold. Intermediate signal levels between the two thresholds suggest that further tests should be made. These thresholds are different for different kind of samples, and their values are obtained according to the reference data, but an experienced person may change the default values [18]. The signal produced by tested samples may vary according to their mineral content [8]. It depends on the quantity and type of inorganic material in food. Also, it may weaken after exposure of samples to natural or artificial light [18–19]. This phenomenon, usually called optical fading, is observed especially during first six months of exposure, during which the luminescence signal drops down continuously [9]. Studies have shown that natural light has a greater effect on optical fading than artificial light [20]. A high signal may be detected even with samples of food that have not been irradiated. The reason for this is their natural exposure to ionizing radiation from winds, soil, *etc.* [21] When using an instrument in a conjunction with a computer, the SUERC PPSL Console is necessary. It presents graphically the dependence of the PSL signal counts on time, by measuring the counts once in a second for 60 seconds. It also calculates the standard deviations from these results. When using it in the stand-alone mode, the intensity of measured signal is indicated by three LED indicators. The green indicator is turned on when the produced signal is below the low threshold ("negative" result). The yellow indicator shows that the signal is between the two thresholds ("intermediate" result). The red indicator is turned on when the signal is above the high threshold ("positive" result).

### *2.2. Samples*

The tested samples of herbs and spices including: paprika, alfalfa, dong qui, green tea, mint, turmeric and thyme tea, have been received from the Scottish Universities Research and Reactor Centre. For the measurements, the samples were placed in disposable plastic Petri dishes suitable for the instrument, in the form of a thin layer. The tests were performed on two portions of each sample. As the measurement lasted for 60 seconds, and the results were obtained once in a second, 60 points were presented graphically for each sample, giving the dependence of the total counts on the dose. Irradiation of the samples was done using Co-60 as an irradiation source at the Institute of Nuclear Sciences in Vinca, Serbia. In order to obtain doses of 1 kGy, 5 kGy and 10 kGy, the irradiation time was 204 seconds, 17 minutes and 34 minutes, respectively.

### *2.3. Description of measurements*

The measurements were performed according to the procedure described in [22]. The detection of irradiated food using PSL were performed by two methods: the screening method and the calibrated method. The screening method does not need any special preparation of samples. The calibrated method is used for validation of the results obtained with the screening method. For performing the calibrated method, the samples are exposed to ionizing radiation after the initial screening measurement. After that, the measurement of an irradiated sample is repeated. If the sample has been originally irradiated, only a small rise in the PSL signal counts after this exposure to ionizing radiation will be observed, while unirradiated samples usually show a significant increase in the PSL signal counts. The samples were exposed to a dose of 1 kGy for performing the calibrated method [18]. The obtained results are presented graphically and eventually a certificate is prepared. The procedure is confirmed by applying

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 143–151.

it to a number of samples. The calibrated method is also used when the samples have been exposed to light, so that the screening method does not give reliable results. The calibrated method is used for precise distinguishing the unirradiated samples and samples that are not sensitive to PSL. It is recommended that the measurements are undertaken for two portions of each sample. In the case the results obtained for these portions are not conclusive, the measurements should be performed for four more portions of the sample, and the two highest results should be taken into consideration. In this paper we performed measurements on seven different samples of herbs and spices. All samples were primary tested by the screening method. The samples were also tested by using the calibrated method, after exposing each sample to a dose of 1 kGy using Co-60. The calibrated method of PSL was performed in order to validate and confirm the results. The tests were also performed after exposition of samples to doses of 5 kGy and 10 kGy. The absorbed dose was controlled using the ethanol chlorine benzene standard, with an uncertainty of 2%. The irradiation system is calibrated at RISO Laboratory in Denmark.

## 3. Results and discussion

The obtained results for PSL measurements on all tested samples are shown in Table 1. The following figures present graphically the dependence of the PSL signal counts on time. The detection system measures the counts once in a second for 60 seconds, so 60 points are presented on each chart. The software calculates the standard deviations from these results. The samples give different results after the same treatment because of their different mineral content.

Table 1. The registered total counts *N*, from the PSL measurements on the samples.

| No. | Sample | *N* | | | |
|---|---|---|---|---|---|
| | | 0 kGy | 1 kGy | 5 kGy | 10 kGy |
| 1 | Paprika standard | 1261 ± 48 | 946501±973 | 2243722 ± 1498 | 2773709 ± 1666 |
| 2 | Alfalfa | 1988 ± 60 | 873252 ± 935 | 2718305 ± 1649 | 3131849 ± 1770 |
| 3 | Dong qui | 2051 ± 56 | 1186266 ± 1090 | 2330155 ± 1527 | 2838005 ± 1685 |
| 4 | Green tea | 739 ± 43 | 26733 ± 167 | 76220 ± 278 | 106669 ± 328 |
| 5 | Mint | 485 ± 40 | 78981 ± 283 | 154225 ± 394 | 206810 ± 456 |
| 6 | Turmeric | 513 ± 40 | 104990 ± 326 | 353964 ± 596 | 355481 ± 597 |
| 7 | Thyme tea | 897 ± 44 | 55419 ± 238 | 112956 ± 338 | 163970 ± 406 |

The total counts ln *N*, as a function of time *t*, for some of the tested samples, are presented in Fig. 3 and Fig. 4.

A paprika standard sample gave an intermediate screening result, which does not give an exact identification of the irradiation history of the sample. Because of that, measurements using the calibrated method were performed after treating the samples by ionizing radiation with a dose of 1 kGy using Co-60 as a source of ionizing radiation. After the treatment, measurements were done for the second time and a positive result was obtained. This confirms that the sample was not previously irradiated. After treatment with higher doses, an increase of the number of total counts has been observed. The above observation is presented in Fig. 3. The alfalfa, paprika and dong qui samples showed similar results.

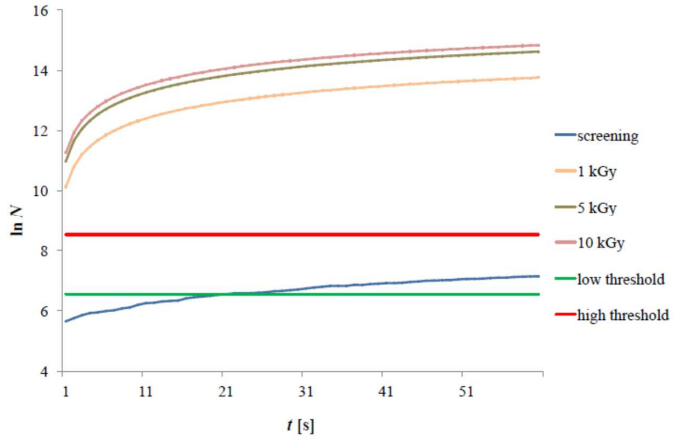



Fig. 3. The PSL measurements of paprika standard for an unirradiated sample (screening method) and samples treated with doses of 1 kGy, 5 kGy and 10 kGy (Co-60 as an irradiation source).

A mint sample gave a negative screening result. After treatment with a dose of 1 kGy, a positive result was obtained, which confirms that the sample has not been irradiated before testing. After treatment with higher doses, an increase of the number of total counts has been noticed. The thyme tea sample gave similar results. The results for a mint sample are shown in Fig. 4. Similar results were obtained for the samples of green tea, mint, turmeric and thyme tea.

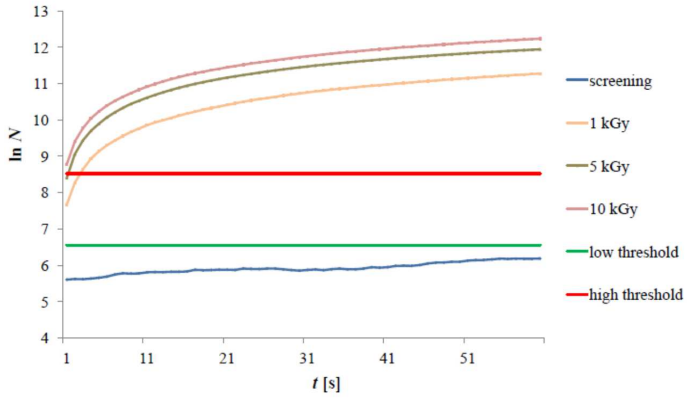



Fig. 4. The PSL measurements of mint for an unirradiated sample (screening method) and samples treated with doses of 1 kGy, 5 kGy and 10 kGy (Co-60 as an irradiation source).

The obtained results confirm validity of the procedure for all tested samples. All of them are correctly identified as unirradiated. Besides that, the study of the influence of dose on the luminescence signal in all cases showed that the higher the dose applied on samples, the greater the luminescence signal. This can serve as a base for further studies of dose estimation [9–10].

The results of PSL measurements for all tested samples irradiated with different doses are shown in Table 2. Standard deviation – $\sigma$ and measurement uncertainty – $u$ for the total counts from all 60 values have been calculated according to the following formulas:

$$\sigma = \sqrt{\frac{\sum_i (\ln N_i - \overline{\ln N})^2}{59}}, \tag{1}$$

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 143–151.

and

$$u = \frac{\sigma}{\sqrt{60}}. \tag{2}$$

Table 2. Standard deviation – $\sigma$ and measurement uncertainty – $u$ of total counts from the PSL measurements on samples.

| | Sample | ln *N* | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 kGy | | 5 kGy | | 10 kGy | |
| | | *σ* | *u* | *σ* | *u* | *σ* | *u* |
| 1 | Paprika standard | 0,21239038 | 0,02741948 | 0,238821415 | 0,030831712 | 0,231563694 | 0,029894744 |
| 2 | Alfalfa | 0,24378165 | 0,03147208 | 0,29664052 | 0,03829613 | 0,29372451 | 0,03791967 |
| 3 | Dong qui | 0,24109032 | 0,03112463 | 0,27621648 | 0,03565939 | 0,29031491 | 0,03747949 |
| 4 | Green tea | 0,21229289 | 0,02740689 | 0,25483323 | 0,03289883 | 0,31458119 | 0,04061226 |
| 5 | Mint | 0,1962257 | 0,02533263 | 0,2613148 | 0,0337356 | 0,24707366 | 0,03189707 |
| 6 | Turmeric | 0,22113866 | 0,02854888 | 0,32828703 | 0,04238167 | 0,32247145 | 0,04163089 |
| 7 | Thyme | 0,275362 | 0,03554908 | 0,21316774 | 0,02751984 | 0,20445422 | 0,02639493 |

The values presented in Table 2 show that there are no significant differences between the deviations for different doses. The deviations are very small, which means that the measurements have been performed with a very low measurement uncertainty.

The dependence of total counts ln $N$ on dose $D$ for all tested samples is shown in Fig. 5 and Fig. 6.

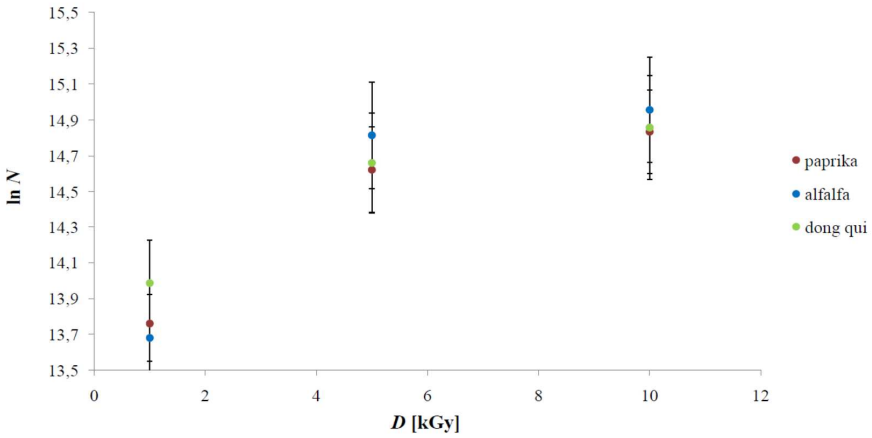



Fig. 5. A plot of total counts against dose for the paprika, alfalfa and dong qui samples.

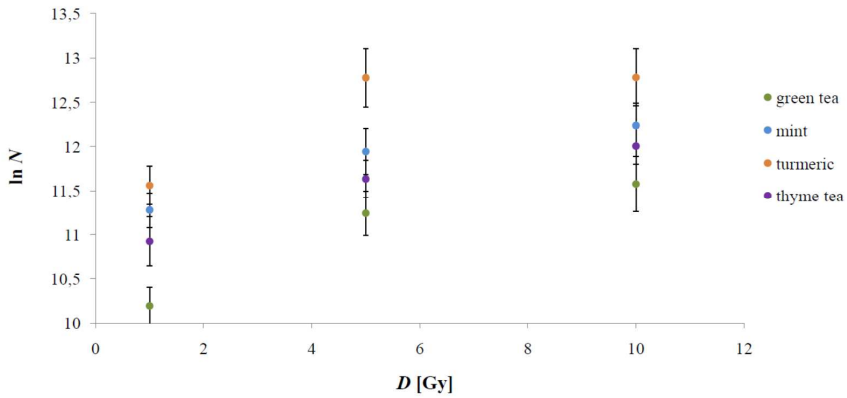



Fig. 6. A plot of total counts against dose for the green tea, mint, turmeric and thyme tea samples.

## 4. Conclusion

Irradiation enhances food quality by destroying harmful microorganisms.

The aim of the presented research was to study the dependence of the luminescence signal on the applied radiation dose. Thus, a number of measurements and analysis of different samples have been undertaken, indicating that the results are comparable with the EU standard, and validity of the procedure for analysis of irradiated food is completely confirmed. Also, a study of the influence of radiation doses on the luminescence signal has been undertaken, which showed that there is a dependence of the signal on the radiation dose.

Even though dosimeters are used for the dose determination during treatment of foodstuffs, the exact absorbed dose may vary depending on the geometry of the package and quantity of packed food. According to the results of the work presented in this paper, which confirm that there is a clear dependence of the luminescence signal on a dose, we suggest that the PSL measurements can serve as an initial step to establishing a procedure of the dose estimation for different types of irradiated food. An exact determination, or even estimation, of the absorbed dose cannot be done regardless the storage conditions, mineral content, irradiation conditions, *etc*. The results also show that the deviations of the results are not significant and that the measurements were performed with a low measurement uncertainty, confirming the reliability of the measurement procedure.

## Acknowledgements

The authors express their strong gratitude to IAEA for funding the TC project – MAK 5007 “Assessing and Enabling the Implementation of Food Irradiation Technologies” and to the Institute of Nuclear Sciences in Vinca, Serbia, for irradiation of the samples.

## References

[1] Miller, R.B. (2005). *Electronic irradiation of foods*. Springer.

[2] Padua, W.G. *Food Irradiation*. Department of Food Science and Human Nutrition.

[3] Meier, W. (1991). Analysis of irradiated food. *Mikrochimica Acta*, 104(1), 71–79.

[4] Manzoli, J.E., Rosa, F.M.L., Silva Felix, J., Monteiro, M. (2007). Migration measurements from plastic packaging: a simulation study on influence of initial concentration profile. *Metrol. Meas. Syst.*, 14(1), 117–124.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 143–151.

[5] Goulas, A.E., Stahl, M., Riganakos, K.A. (2008). Effect of various parameters on detection of irradiated fish and oregano using the ESR and PSL methods. *Food control*, 19(11), 1076–1085.

[6] Analytical methods for post-irradiation dosimetry of foods – Technical report. (1993). IUPAC.

[7] Cutrubinis, M., Delincee, H., Stahl, M., Roder, O., Schaller, H. J. (2005). Detection methods for cereal grains treated with low and high energy electrons. *Radiation Physics and Chemistry*, 72(5), 639–644.

[8] Soika, C., Delincee, H. (2000). Thermoluminescence Analysis for Detection of Irradiated Food – Luminescence Characteristics of Minerals for Different Types of Radiation and Radiation Doses. *Academic Press*, 33, 431–439.

[9] Ahn, J., Kim, G., Akram, K., Kim, K., Kwon, J. (2012). Luminescence characteristics of minerals separated from irradiated onions during storage under different light conditions. *Radiation Physics and Chemistry*, 81(8), 1215–1219.

[10] D'Oca, M.C., Bartolotta, A., Cammilleri, C., Giuffrida, S., Parlato, A., Di Stefano, V. (2009). The additive dose method for dose estimation in irradiated oregano by thermoluminescence technique. *Food control*, 20(3), 304–306.

[11] Kim, B.K., Akram, K., Kim, C.T., Kang, N.R., Lee, J. W., Ryang, J.H., Kwon, J.H. (2012). Identification of low amount of irradiated spices (red pepper, garlic, ginger powder) with luminescence analysis. *Radiation Physics and Chemistry*, 81(8), 1220–1223.

[12] Directive 1999/2/EC of the European Parliament and of the Council, 1999.

[13] Directive 1999/3/EC of the European Parliament and of the Council, 1999.

[14] Law of food safety and products and materials that come in contact with food, of the Statute of Republic of Macedonia (Official Gazette of Republic of Macedonia, No. 54/2002 and 84/2007).

[15] Regulation for specific security requirements of the food produced by ionizing radiation (Official Gazette of Republic of Macedonia, No. 63/2014).

[16] The SUERC Pulsed Photostimulated Luminescence Irradiated Food Screening System – User Manual.

[17] MKS EN 13751:2011 Foodstuffs – Detection of irradiated food using photostimulated luminescence.

[18] Bortolin, E., Boniglia, C., Calicchia, A., Alberti, A., Fuochi, P., Onori, S. (2007). Irradiated herbs and spices detection: light-induced fading of the photo-stimulated luminescence response. *International Journal of Food Science and Technology*, 42(3), 330–335.

[19] Bayram, G., Delincée, H. (2004). Identification of irradiated Turkish foodstuffs combining various physical detection methods. *Food control*, 15(2), 81–91.

[20] Ahn, J., Kim, G., Akram, K., Kim, K., Kwon, J. (2012). Effect of storage conditions on photostimulated luminescence of irradiated garlic and potatoes. *Food Research International*, 47(3), 315–320.

[21] Jo, D., Kim, B., Kausar, T., Kwon, J. (2008). Study of photostimulated- and thermo-luminescence characteristics for detecting irradiated kiwifruit. *Journal of Agricultural and Food Chemistry*, 56(4), 1180–1183.

[22] Ginovska, M., Spasevska H., Stojanovska-Georgievska L., Sandeva, I., Kochubovski, M. (2016). Procedure for detection and control of irradiated food. *Journal of Environmental Protection and Ecology*, 17(1), 402–412.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 153–160.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# EXAMINATION OF SOL-GEL DERIVED HYDROXYAPATITE ENHANCED WITH SILVER NANOPARTICLES USING OCT AND RAMAN SPECTROSCOPY

**Maciej J. Głowacki[1,3], Marcin Gnyba[1], Paulina Strąkowska[1,3], Mateusz Gardas[2], Maciej Kraszewski[1], Michał Trojanowski[1], Marcin R. Strąkowski[1]**

*1) Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics, G. Narutowicza 11/12, 80-233 Gdańsk, Poland (margnyba@pg.gda.pl, pauanton@pg.gda.pl, mackrasz@student.pg.gda.pl, mictroja@pg.gda.pl*
*✉ marcin.strakowski@eti.pg.gda.pl, +48 58 347 1361)*
*2) Mateusz Gardas GAROCIN LABS, Świerkowa 4, 76-200 Słupsk, Poland (gardas.mateusz@gmail.com)*
*3) Gdańsk University of Technology, Faculty of Mechanical Engineering, G. Narutowicza 11/12, 80-233 Gdańsk, Poland (macglow1@student.pg.gda.pl)*

**Abstract**

*Hydroxyapatite* (HAp) has been attracting widespread interest in medical applications. In a form of coating, it enables to create a durable bond between an implant and surrounding bone tissues. With addition of silver nanoparticles HAp should also provide antibacterial activity. The aim of this research was to evaluate the composition of hydroxyapatite with silver nanoparticles in a non-destructive and non-contact way. For control measurements of HAp molecular composition and solvent evaporation efficiency the Raman spectroscopy has been chosen. In order to evaluate dispersion and concentration of the silver nanoparticles inside the hydroxyapatite matrix, the *optical coherence tomography* (OCT) has been used. Five samples were developed and examined – a reference sample of pure HAp sol and four samples of HAp colloids with different silver nanoparticle solution volume ratios. The Raman spectra for each solution have been obtained and analyzed. Furthermore, a transverse-sectional visualization of every sample has been created and examined by means of OCT.

Keywords: hydroxyapatite, sol-gel, nanoparticles, Raman spectroscopy, optical coherence tomography (OCT).



## 1. Introduction

Hydroxyapatite, $Ca_{10}(PO_4)_6(OH)_2$, commonly referred to as HAp, has been widely applied in bone surgery, implantation and dentistry. Biological (nonstoichiometric) hydroxyapatite is an inorganic compound naturally existing as a structural template for the mineral phase of vertebrate bones and teeth [1–2]. In elevated temperatures, when an appropriate metastable medium is applied, biological hydroxyapatite may also be obtained from marine invertebrate cartilage tissues [3]. Synthetic (stoichiometric) hydroxyapatite is a bio-ceramic material which can be manufactured using various methods, including wet precipitation, sol-gel processing, solid-state reactions, hydrothermal treatment or a flux method [4–10]. The sol-gel technique has several advantages over other methods suitable for preparing synthetic HAp. Firstly, it enables to replace high-temperature reactions of synthesis with low-temperature processes [11]. Secondly, it may be used for creating both fine powders and thin ceramic layers on complex surfaces [12]. Finally, due to a high fluidity of the sols, the samples prepared using the sol-gel processing are irradiated by light in their entire depth and therefore may be examined using optical measurement methods.

Due to its structural and compositional similarities with the natural bone, the artificial hydroxyapatite has the highest biocompatibility among currently known ceramic biomaterials [13–14]. It enables to create a durable biological bonding with surrounding bone tissues, causes osteoinductive activity and has no negative effects on human organism [4]. Unfortunately, HAp

Article history: received on Jul. 27, 2016; accepted on Sep. 24, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0008.

does not exhibit antibacterial activity which would reduce the risk of infection during and after the insertion of an implant. To provide the hydroxyapatite with antiseptic properties, current research has been focused on combining the biomaterial with silver nanoparticles [15–16]. However, uniform distribution of the nanoparticles must be maintained in order to optimize material properties and to avoid disruption of the HAp structure. Poor mechanical properties of hydroxyapatite make it insufficient to be used as an implant experiencing severe stresses. Instead, HAp is often combined with metallic or polymer phases to create composites of satisfactory biomechanical performance.

The aim of this research was to develop a non-destructive optical measurement methodology for analysis of silver-hydroxyapatite nanocomposites prepared by the sol-gel technology. Regardless of the hydroxyapatite manufacturing process, the most common methods for its characterization are *scanning electron microscopy* (SEM), *X-ray diffraction* (XRD) and *Fourier-transform infrared* (FTIR) spectroscopy [5–10]. However, in order to examine an object supposed to work in the environment of human organism, it is reasonable to use techniques which enable to examine the material without changing its inner structure or causing damage. Conventional SEM is generally destructive due to special preparation methods which provide samples with electrical conductivity. Although FTIR spectroscopy and XRD are considered non-invasive, these techniques do not enable to visualize the internal structure of the analysed material. Therefore, in order to evaluate dispersion and concentration of silver nanoparticles inside the hydroxyapatite matrix, the *optical coherence tomography* (OCT) has been chosen. OCT has never been extensively used for HAp inspection and it is one of a few methods suitable for evaluation of nanocomposite materials which belong to the NDE/NDT group. For a qualitative analysis and control measurements of HAp the Raman spectroscopy has been selected.

## 2. Measurement techniques

The *optical coherence tomography* (OCT), which has been used in this study to examine distribution of silver nanoparticles inside the hydroxyapatite colloids, is a contact-free, non-invasive measurement technique widely known for its biomedical applications, especially in ophthalmology [17]. It has also proven to be useful in evaluation of technical objects and recent research results have shown its potential for nanocomposite material inspection [18–19]. OCT is based on low-coherence interferometry, therefore the light backscattered from particular points inside the evaluated material can be spatially detected and recorded. As a result, the depth-resolved reflectivity profiles of the sample, called A-scans, are obtained. Combining series of laterally *adjacent depth-scans* (A-scans) enables to generate cross-sectional images of the samples, known as B-scans. The OCT technique also enables to create 3D tomographic pictures. Although standard OCT systems deliver only intensity images, the backscattered signal provides more valuable data, *i.e.* spectral characteristic or polarization state of the light [20]. This additional information can be specified in the pictures by various image contrasting algorithms. Thereby, OCT may be extended to more advanced and sophisticated systems like *polarization-sensitive optical coherence tomography* (PS-OCT), which has been used in this study and is presented in Fig. 1 [20–21].

In the PS-OCT system, a light beam emitted by a broadband swept source is used for the measurements. The beam is linearly polarized before entering a *beam-splitter* (BS), which divides the incoming light into two sections - a sample arm containing the evaluated material and a reference arm, which includes a mirror. In the presented solution a Michelson interferometer is used. However, other types of two-beam interferometers, like Mach-Zehnder, can be successfully applied. Both arms of the interferometer contain *quarter-wave plates* (QWP), which let through all the incoming light, providing a stable polarization control, which

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 153–160.

is necessary for polarization sensitive analysis. The beam backscattered by the sample interferes with the light reflected from the reference arm at the *beam-splitter* (BS). Then, the *polarizing beam-splitter* (PBS) separates two orthogonal polarization states – horizontal and vertical – in order to provide detection of polarization diversity. Such a polarization-sensitive system delivers information about local changes of birefringence inside the sample structure.

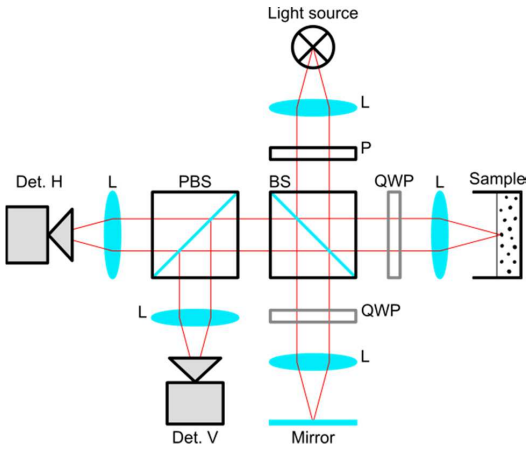



Fig. 1. A schematic diagram of a polarization-sensitive optical coherence tomography system. L – lens; P – polarizer; BS – *beam-splitter*; QWP – *quarter-wave plate*; PBS – *polarizing beam-splitter*; Det. – detector: V – vertical, H – horizontal.

A complementary method used to characterize the samples according to their optical properties is the Raman spectroscopy, which is a contact-free and non-invasive measurement technique based on inelastic scattering of light and is used for molecular identification of materials and their quantitative analysis [22–23]. The method, which is also complementary to IR spectroscopy, may be used to study solid, liquid and gaseous substances. The Raman spectroscopic system which has been used in this research is presented in Fig. 2.

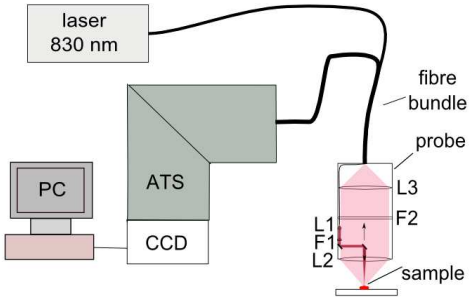



Fig. 2. A schematic diagram of the Raman spectroscopic system. L1, L2, L3 – lenses; F1 – laser line filter; F2 – low-pass filter; ATS – axial transmissive spectrograph; CCD – detector array; PC – computer.

The excitation wave signal from the laser can be delivered either through an open space or by a fiber optic probe. The laser line filter can be applied to ensure irradiation of samples by a single-frequency light. The collected scattering signal is transmitted to the spectrograph. When the Raman scattering occurs, the registered spectrum of light contains not only the

Rayleigh band with a frequency identical to that of the incident light but also symmetrically distant bands with a decreased and increased frequency, called Stokes and anti-Stokes lines, respectively. Their number and location depend on the internal structure of the samples and are unique for every material. Since the Raman scattering is very weak, there is a number of procedures taken to distinguish it from the predominant Rayleigh scattering, *e.g.* implementation of a low-pass or notch filter in the collection part. Moreover, the detector array is cooled to improve the signal-noise ratio. The Raman spectroscopy is non-destructive for the samples as long as the radiation intensity is controlled.

## 3. Experimental procedure

In the presented experiment, the silver-hydroxyapatite composite was synthesised by mixing an aqueous solution of silver nanoparticles with the hydroxyapatite colloid prepared using the sol-gel technique.

Calcium nitrate tetra-hydrate $Ca(NO_3)_2$ x $4H_2O$ and di-phosphorus pentoxide $P_2O_5$ were selected for the synthesis of HAp by the sol-gel method [24]. First, the di-phosphorus pentoxide was dissolved in ethyl alcohol, then the calcium nitrate tetra-hydrate was added to the solution to attain a Ca/P molar ratio of 1.67, which is the necessary condition for obtaining the synthetic hydroxyapatite. The transparent alcoholic solution of $P_2O_5$ turned into an opaque colloid after the addition of $Ca(NO_3)_2$ x $4H_2O$. A sample of pure HAp sol was prepared in a Petri dish and examined by the OCT method before the aging process. The Raman spectrum of the sample was also obtained and analysed.

In order to prepare the solution of silver nanoparticles, silver nitrate $AgNO_3$ and ascorbic acid were separately dissolved in deionized water containing polyvinyl alcohol PVA [25]. Then the aqueous solution of the ascorbic acid was added dropwise into the silver nitrate solution, which made the transparent liquid turn red. The resulting colloid, which darkened after being kept statically for 30 min, was added to the hydroxyapatite sol and stirred energetically. Four sols were prepared with different HAp sol to Ag nanoparticle solution volume ratios: 1/2, 1/1, 2/1, 3/1. A sample of each sol was prepared in the Petri dish and examined by means of OCT and the Raman spectroscopy. The diameter of silver nanoparticles was estimated to be below 300 nm.

The PS-OCT system used in this study has been developed at the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology (Poland). The most important features of the OCT measurement system are summarized in Table 1.

Table 1. The PS-OCT system features.

| Item | Value |
|---|---|
| Light source type | 20 kHz swept source (SS) |
| Average output power | 10 mW |
| Central wavelength | 1320 nm |
| Wavelength range | 140 nm |
| Axial resolution | 12 μm |
| Lateral resolution | 15 μm |
| Frame rate | > 4 fps |
| Max. displaying imaging range / transverse imaging range | 7 mm / 10 mm |

The Raman spectroscopy was performed in a system based on a pre-commercial Ramstas spectrometer developed by the VTT – Technical Research Centre of Finland [26–28]. The characteristics of the Raman spectrometer system used in this research are presented in Table 2.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 153–160.

Table 2. The Raman spectrometer characteristics.

| Item | Value |
|---|---|
| Light source type | Diode laser / CW mode, central wavelength 830 nm |
| Average power on sample | 100 mW |
| Spectrograph | Axial transmissive setup with holographic transmission grating |
| Detector | TE-cooled CCD array / 1024 rows |
| Spectral range | 200 – 2000 $cm^{-1}$ |
| Spectral resolution | 8 $cm^{-1}$ |
| Optical system | Fibre optics probe / working distance – 5 cm |

## 4. Results and discussion

The samples of hydroxyapatite placed in the Petri dishes were successfully measured using the polarization-sensitive optical coherence tomography. The obtained cross-sectional visualizations of every sample were examined. The resulting OCT images, which are presented in Fig. 3 with corresponding photographs of HAp samples, were well-detailed at any depth of the analyzed material.

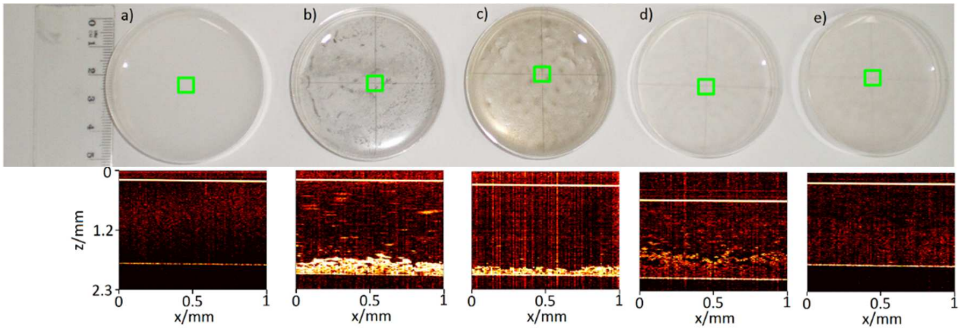



Fig. 3. The OCT images with corresponding photographs of HAp samples.
Pure Hap (a); HAp/nanoparticles = 1/2 (b); HAp/nanoparticles = 1/1 (c); HAp/nanoparticles = 2/1 (d); HAp/nanoparticles = 3/1 (e). Green squares in the photographs indicate the OCT measurement areas (5.5 mm by 5.5 mm).

The OCT picture of pure HAp sol was used as a reference and every other sample was compared with it in order to evaluate the distribution of the silver nanoparticles inside the hydroxyapatite matrix. Although very small diameters of the nanoparticles place them below the resolution of the OCT system, and therefore their real dimensions could not be assessed in this measurement, still they are clearly visible in the pictures against the hydroxyapatite background and their degree of dispersion may be estimated. The sample with the lowest Ag content (the HAp sol to Ag nanoparticle solution volume ratio equal to 3/1) bears the closest resemblance to the reference sample of pure hydroxyapatite sol. This indicates that the nanoparticles are uniformly distributed inside the hydroxyapatite and the structure of the composite is quite homogeneous. In the case of samples with higher concentrations of silver, the nanoparticles have sunk to the bottom of Petri dishes and agglomerated into new phases, which are clearly visible in the images.

The Raman spectra of the sols obtained after mixing the components as well as the spectra of gels recorded after 24 hours of drying are shown in Fig. 4. Main bands at 880 $cm^{-1}$ and 1045 $cm^{-1}$ can be assigned to ethanol and ν(P-O) stretching mode of HAp, respectively.

Comparison of their intensity shows how efficiently ethanol evaporates during the drying process and how the structure of HAp is being created. In the case of spectra of the pure HAp sol and the sample with the lowest concentration of Ag nanoparticles the gelation was quite effective – the ethanol mostly evaporated during 24 hours and the (P-O)-based network of HAp was developing. This stays in a good agreement with the results obtained by OCT, which have shown that these samples present the best homogeneity and distribution of the nanoparticles. Contrary to them, samples with a higher Ag content still contained a larger content of non-vaporised ethanol and agglomerated Ag particles after 24 hours. It suggests that components of Ag colloids may interrupt ethanol evaporation and thus disturb uniform distribution of the silver nanoparticles. Moreover, these samples contain a huge amount of different unreacted particles and clusters which produces a stronger optical background.

The band assigned to P-O stretching is shifted towards higher wavenumbers in comparison to crystalline HAp (962 $cm^{-1}$) because materials were still in an elastic or liquid form (some ethanol remained inside) during the examination [29].

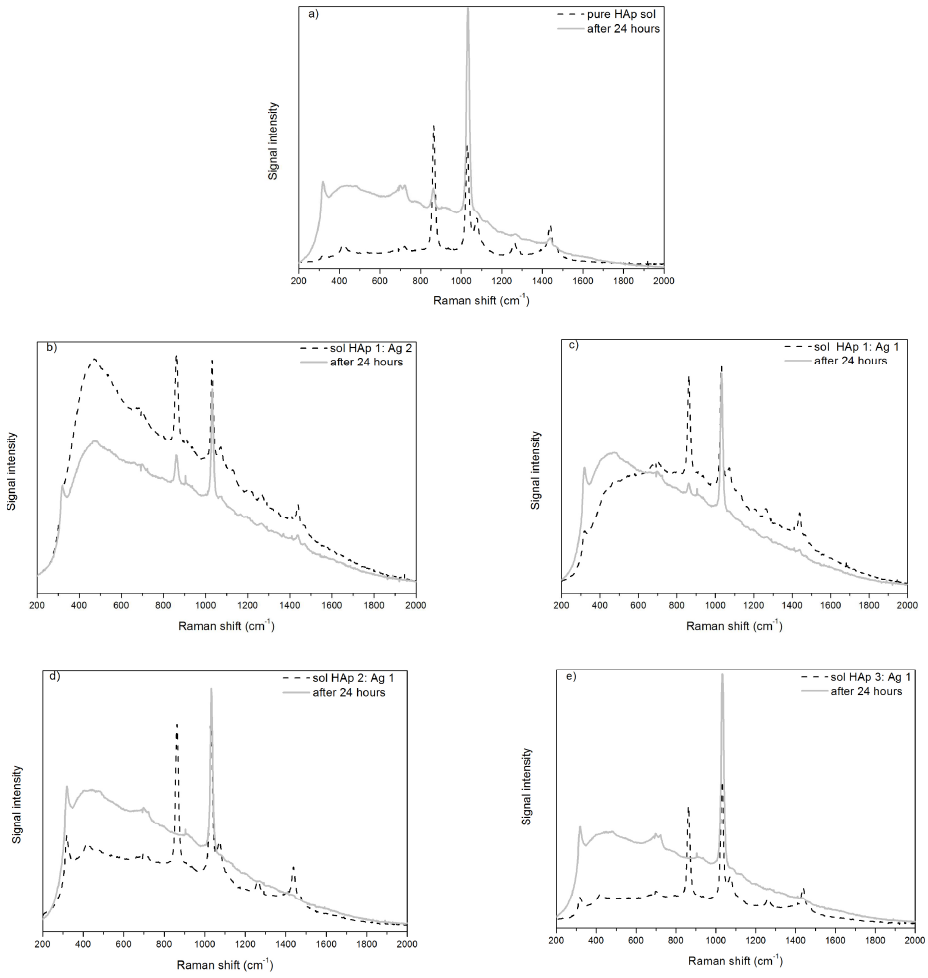



Fig. 4. The Raman spectra of HAp samples. Pure Hap (a); HAp/nanoparticles = 1/2 (b); HAp/nanoparticles = 1/1 (c); HAp/nanoparticles = 2/1 (d); HAp/nanoparticles = 3/1 (d). The spectra were recorded after synthesis of the sols and after 24 hours of drying.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 153–160.

## 5. Conclusions

The results of presented research have confirmed a significant potential of simultaneous use of the optical coherence tomography and the Raman spectroscopy for non-destructive examination of nanocomposite materials based on HAp with Ag nanoparticles. Their use has enabled to select a sample with the best distribution of silver nanoparticles (in this particular case – the sample with the ratio of HAp/nanoparticles = 3/1) as well as to study processes which take place inside the material during the gelation and the lattice formation. The differences between light scattering determined by the OCT can be correlated with the differences of molecular composition of the material examined by the Raman spectroscopy. We correlated the better homogeneity of the sample and distribution of the Ag nanoparticles (OCT profile analysis) with the most efficient ethanol evaporation during the drying and with consequent better lattice formation (change of the intensity of respective Raman bands). Thus, our results have shown that the addition of the aqueous solution of the silver nanoparticles to the HAp sol may heavily disrupt the gelation and drying processes of the composite if the content of Ag solution is too high. The reason of such a disorder could be the fact that the two solutions were based on different solvents. Preparation of an aqueous HAp sol or an alcoholic solution of the silver nanoparticles could possibly solve the problem if introduction of a higher content to HAp is required.

## Acknowledgements

This research work has been supported by The National Centre for Research and Development (NCBiR), Poland, under the grant no. LIDER/32/205/L-3/11 and the DS program of Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology. The authors wish to thank the researchers from VTT – Technical Research Centre of Finland (Oulu) for providing a VTT Ramstas spectrometer.

## References

[1] Leventouri, Th. (2006). Synthetic and biological hydroxyapatites: Crystal structure questions. *Biomaterials*, 27, 3339–3342.

[2] Sadat-Shojai, M., Khorasani, M.T., Dinpanah-Khoshdargi, E., Jamshidi, A. (2013). Synthesis methods for nanosized hydroxyapatite with diverse structures. *Acta Biomaterialia*, 9, 7591–7621.

[3] Eilberg, R.G., Zuckerberg, D.A. (1975). Mineralization of Invertebrate Cartilage. *Calcif. Tiss. Res.*, 19, 85–90.

[4] Orlovskii, V.P., Komlev, V.S., Barinov, S.M. (2002). Hydroxyapatite and Hydroxyapatite-Based Ceramics. *Inorganic Materials*, 38(10), 973–984.

[5] Afshar, A., Ghorbani, M., Ehsani, N., Saeri, M.R., Sorrell, C.C. (2003). Some important factors in the wet precipitation process of hydroxyapatite. *Materials and Design*, 24, 197–202.

[6] Mobasherpour, I., Soulati Heshajin, M., Kazemzadeh, A., Zakeri, M. (2007). Synthesis of nanocrystalline hydroxyapatite by using precipitation method. *Journal of Alloys and Compounds*, 430, 330–333.

[7] Bogdanoviciene, I., Beganskiene, A., Tonsuaadu, K., Glaser, J., Meyer, H.J., Kareiva, A. (2006). Calcium hydroxyapatite, $Ca_{10}(PO_4)_6(OH)_2$ ceramics prepared by aqueous sol-gel processing. *Materials Research Bulletin*, 41, 1754–1762.

[8] Pramanik, S., Agarwal, A.K., Rai, K.N., Garg, A. (2007). Development of high strength hydroxyapatite by solid-state-sintering process. *Ceramics International*, 33, 419–426.

[9] Wang, Y., Zhang , S., Wei, K., Zhao, N., Chen, J., Wang, X. (2006). Hydrothermal synthesis of hydroxyapatite nanopowders using cationic surfactant as a template. *Materials Letters*, 60, 1484–1487.

[10] Teshima, K., Lee, S., Sakurai, M., Kameno, Y., Yubuta, K., Suzuki, T., Shishido, T., Endo, M., Oishi, S. (2009). Well-Formed One-Dimensional Hydroxyapatite Crystals Grown by an Environmentally Friendly Flux Method. *Crystals Growth & Design*, 9(6), 2937–2940.

[11] Mohseni, E., Zalnezhad, E., Bushroa, A.R. (2014). Comparative investigation on the adhesion of hydroxyapatite coating on Ti-6Al-4V implant: A review paper. *International Journal of Adhesion & Adhesives*, 48, 238–257.

[12] Brinker, C.J., Scherer, G.W. (1990). *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing*. USA: Academic Press.

[13] Uklejewski, R., Winiecki, M., Mielniczuk, J., Rogala, P., Auguściński, A. (2008). The poroaccessibility parameters for three-dimensional characterization of orthopedic implants porous coatings. *Metrol. Meas. Syst.*, 15(2), 215–226.

[14] Batory, D., Gawronski, J., Kaczorowski, W., Niedzielska, A. (2012). C-HAp composite layers deposited onto AISI 316L austenitic steel. *Surface & Coatings Technology*, 206, 2110–2114.

[15] Andrade, F.A.C., de Oliveira Vercik, L.C., Monteiro, F.J., da Silva Rigo, E.C. (2016). Preparation, characterization and antibacterial properties of silver nanoparticles-hydroxyapatite composites by a simple and eco-friendly method. *Ceramics International*, 42, 2271–2280.

[16] Tian, B., Chen, W., Yu, D., Lei, Y., Ke, Q., Guo, Y., Zhu, Z. (2016). Fabrication of silver nanoparticle-doped hydroxyapatite coatings with oriented block arrays for enhancing bactericidal effect and osteoinductivity. *Journal of the mechanical behavior of biomedical materials*, 61, 345–359.

[17] Fercher, A.F., Drexler, W., Hitzenberger, C.K., Lasser, T. (2003). Optical coherence tomography - principles and applications. *Reports on Progress in Physics*, 66, 239–303.

[18] Strąkowski, M.R., Pluciński, J., Jędrzejewska-Szczerska, M., Hypszer, R., Maciejewski, M., Kosmowski, B.B. (2008). Polarization sensitive optical coherence tomography for technical materials investigation. *Sensors and Actuators A*, 142, 104–110.

[19] Trojanowski, M., Kraszewski, M., Strąkowski, M.R., Pluciński, J. (2015). Optical Coherence Tomography for nanoparticles quantitative characterization. *Proc. SPIE 9554, Nanoimaging and Nanospectroscopy III*, 95540I.

[20] Strąkowski, M.R., Pluciński, J., Kosmowski, B.B. (2011). Polarization Sensitive Optical Coherence Tomography with Spectroscopic Analysis. *Acta Physica Polonica A*, 120(4), 785–788.

[21] Pircher, M., Hitzenberger, C.K., Schmidt-Erfurth, U. (2011). Polarization sensitive optical coherence tomography in the human eye. *Progress in Retinal and Eye Research*, 30, 431–451.

[22] Kwiatkowski, A., Gnyba, M., Smulko, J., Wierzba, P. (2010). Algorithms of chemicals detection using Raman spectra. *Metrol. Meas. Syst.*, 17(4), 549–560.

[23] Ferraro, J.R., Nakamoto, K., Brown, C.W. (2003). *Introductory Raman Spectroscopy*. Elsevier.

[24] Kim, I., Kumta, P.N. (2004). Sol-gel synthesis and characterization of nanostructured hydroxyapatite powder. *Materials Science and Engineering B*, 111, 232–236.

[25] Zielinska, A., Skwarek, E., Zaleska, A., Gazda, M., Hupka, J. (2009). Preparation of silver nanoparticles with controlled particle size. *Procedia Chemistry*, 1, 1560–1566.

[26] Niemelä, P., Suhonen, J. (2001). Rugged Fiber-Optic Raman Probe for Process Monitoring Applications. *Applied Spectroscopy*, 55(10), 1337–1340.

[27] Gnyba, M., Keränen, M., Maaninen, A., Suhonen, J., Jędrzejewska-Szczerska, M., Kosmowski, B.B., Wierzba, P. (2005). Raman system for on-line monitoring and optimization of hybrid polymer gelation. *Opto-Electronics Review*, 13(1), 9–17.

[28] Gnyba, M., Keraenen, M. (2003). Optical investigation of molecular structure of sophisticated materials for photonics. *Proc. SPIE*, 5125, 339–344.

[29] Koutsopoulos, S. (2002). Synthesis and characterization of hydroxyapatite crystals: A review study on the analytical methods. *Journal of Biomedical Materials Research*, 62(4), 600–612.

www.czasopisma.pan.pl PAN www.journals.pan.pl
POLSKA AKADEMIA NAUK

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 161–173.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# COMPARISON OF TIME WARPING ALGORITHMS FOR RAIL VEHICLE VELOCITY ESTIMATION IN LOW SPEED SCENARIOS

**Stefan Hensel[1)], Marin B. Marinov[2)]**

*1) University of Applied Sciences Offenburg, Department for Electrical Engineering, Badstraße 24, D-77652 Offenburg, Germany (stefan.hensel@hs-offenburg.de)*

*2) Technical University of Sofia, Faculty of Electronic Engineering and Technologies, Kliment Ohridski Blvd., BG-1756 Sofia, Bulgaria (✉ mbm@tu-sofia.bg, +359 888 865 158)*

**Abstract**

Precise measurement of rail vehicle velocities is an essential prerequisite for the implementation of modern train control systems and the improvement of transportation capacity and logistics. Novel eddy current sensor systems make it possible to estimate velocity by using cross-correlation techniques, which show a decline in precision in areas of high accelerations. This is due to signal distortions within the correlation interval. We propose to overcome these problems by employing algorithms from the field of dynamic programming. In this paper we evaluate the application of correlation optimized warping, an enhanced version of dynamic time warping algorithms, and compare it with the classical algorithm for estimating rail vehicle velocities in areas of high accelerations and decelerations.

Keywords: velocity estimation, cross-correlation, dynamic programming, eddy current sensors.



## 1. Introduction

Reliable and precise measurement of the velocity of a rail vehicle is crucial for the application of modern train disposition systems aimed at increasing the efficiency and thus the amount of goods and persons transported on already existing tracks [1]. Current systems are built upon standard velocity sensors like *Global Navigation Satellite System* (GNSS) receivers or radar systems that face problems when dealing with heavy environment conditions or shadowed areas like rail stations or dense forests [2]. In contrast, an eddy current sensor system enables non-contact measurement of speed and distance of rail vehicles by measuring the magnetic inhomogeneities along the track and utilizes the cross-correlation technique to determine the time shift between two sensor heads mounted within the housing at a set distance from each other [3]. The sensor system works effectively, especially at higher velocities. Nonetheless, this type of sensor encounters difficulties in phases of high deceleration and acceleration as well as in passages with very low speed manoeuvres, *e.g.* when passing over turnouts in railway stations. This paper presents a signal processing approach, based on the so-called warping algorithms, a specific application of the dynamic programming [4] so that these problems at lower velocities can be overcome.

Two types of algorithms are examined: the classical *dynamic time warping* (DTW) algorithm [5] and an adapted variant, the so-called *correlation optimized warping* (COW) algorithm [6]. They are compared with the classical cross-correlation approach, based on a closed-loop correlator [7–9]. Warping algorithms are commonly used for the task of sequence classification, where they are capable of distorting one signal sequence by stretching it so that it is comparable to a class template. This paper makes use of this signal straining, as it is directly proportional to the difference of the two signals determined by cross-correlation. Fig. 1 shows an overview

Article history: received on Oct. 15, 2015; accepted on Sep. 24, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0012.

of the system. The $s_1(t)$ and $s_2(t)$ are the output signals from the two sequentially placed *Eddy Current Sensors* (ECSs). As long as the rail vehicle moves below a certain velocity, *i.e.* when starting or coming to a halt, the speed is determined by means of the two warping algorithms. A velocity threshold determines when the common closed loop correlator or the warping algorithms should be used for velocity estimation. When driving faster, which is the case on open tracks, a *closed loop correlator* (CLC) is employed for estimation.

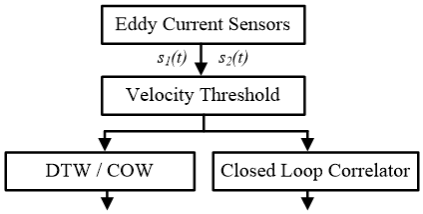



Fig. 1. A system overview.

## 2. Eddy Current Sensor System

### *2.1. Working principle and sensor system*

ECSs are commonly used to detect inhomogeneity in the magnetic resistance of conductive materials [10]. This basic approach has been further developed and adapted for applications on railway vehicles, including speed measurement and pattern recognition tasks [11]. The ECS system consists of two identical sensor devices, each built up with a transmitter coil and two pickup coils. Both sensors are sequentially placed within a housing mounted on the train bogie approximately 10 cm above the railhead. Fig. 2a demonstrates the principle of a single device of the ECS: The transmitter coil $E$ excites a magnetic field $H_E$ that induces eddy currents in metallic materials like the rail. The eddy currents induce an antipode magnetic field $H_{EC}$, that generates $u_{P1}(t)$ and $u_{P2}(t)$ voltages within the $P1$ and $P2$ pick-up coils, respectively. By interconnecting them differentially, the output signal $s(t) = u_{P1}(t) - u_{P2}(t)$ is a measure of rail inhomogeneities. These result mainly from rail clamps, turnouts and other irregularities, *e.g.* cracks or signal cables (see [3]).

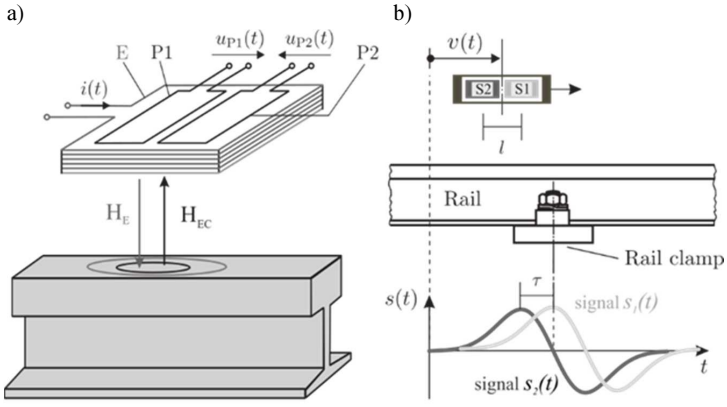



Fig. 2. A single ECS S1 (a); an example of signals in the ECS system (two sensors): $s_1(t)$ and $s_2(t)$, when crossing a rail clamp (b).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 161–173.

The overall signal has a high *signal-to-noise ratio* (SNR), given that pre-processing low pass filters are installed in the sensor hardware.

### *2.2. Correlation based velocity estimation*

The eddy current sensor system generates two signals: $s_1(t)$ and $s_2(t)$, each measured by the sensor heads $S1$ and $S2$, as shown in Fig. 2b and described in the previous section. If the rail vehicle is moving with constant velocity, the resulting signals are actually a low pass filtered sinusoidal of constant frequency and phase shift. This is due to the fact that the equidistance positioned rail clamps induce the main signal part. It is sufficient to know the coil distance $l$ and time shift $T$ to estimate the current velocity with:

$$\mathrm{v} = l/T. \tag{1}$$

The system setup for velocity estimation is shown in Fig. .

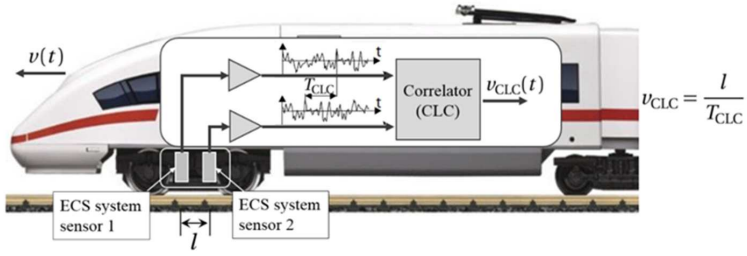



Fig. 3. The system setup for correlation-based velocity measurement.

The time shift within interval $T_f$ is hereby determined with the *cross-correlation function* (CCF) $R_{s_1s_2}(\tau)$ defined as:

$$R_{s_1s_2}(\tau) = \lim_{T_f \to \infty} \frac{1}{T_f} \int_0^{T_f} s_1(t-\tau)\, s_2(t) dt. \tag{2}$$

The CCF has its main maximum at point $\tau = T_{\max}$, , $T_{\max} = \arg\max_{\tau}\{R_{s_1s_2}(\tau)\}$, where the signals are most similar to each other. The main idea now is to determine the time shift $T$ on the basis of this maximum. Since the signals resulting from driving over rail clamps correspond to periodic signals, the CCF is also a periodic function [12]. Thus, it is more complicated to determine $\tau$, as the maximum and its side maxima are indistinguishable in ideal circumstances. To prevent leaping between several maxima, the ECS uses a *Closed-Loop-Correlator* (CLC) that contains a model time-shift to track the peak of the CCF.

CLC is suitable for rail vehicle velocity measurement because of its good dynamic properties, its low statistical error and large measuring range [3]. The hardware implementation is based on the polarity correlation function:

$$R_{12}(\tau) = E\{\mathrm{sgn}[s_1(t)]\, \mathrm{sgn}[s_2(t-\tau)]\}. \tag{3}$$

This enables calculation based solely on the algebraic signs of the signal and significantly reduces the hardware implementation and computational effort.

The corresponding time-shift $T_{CLC}$ of the polarity correlation function lies at its maximum, which is found by optimization with a gradient descent method. This optimization is implemented with the Newton-Raphson algorithm, an iterative method with fast convergence

[13]. The resulting block diagram of the CLC is shown in Fig. 4. Here $g_P(t,\tau)$ and $g_M(t,\tau)$ are the pulse responses of the system and of the model; sgn[...] is the signum function; $e(t)$ – the error and $\tau$ – the model run-time.

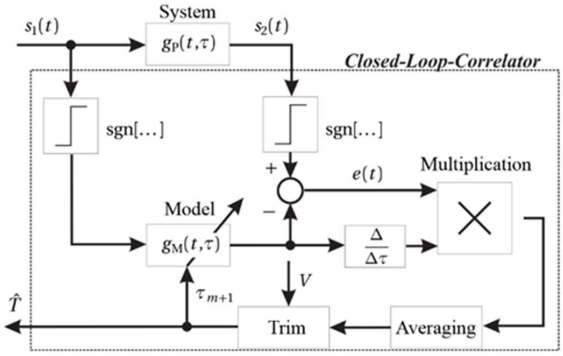



Fig. 4. The working principle of the described CLC. It is created as a signum correlator, optimized with an iterative Newton-Raphson scheme.

## 3. Velocity estimation with time warping algorithms

The above-mentioned approaches base on the assumption of a stationary stochastic process, which is true for a constant velocity within the cross-correlation interval. Although this assumption is correct in most situations, it is heavily violated in low-speed manoeuvres, where large changes in the relative velocity may occur. Unfortunately, this is the case in safety-relevant areas, *e.g.* within stations, where there are many turnouts and they additionally disturb the signals. The need for reliable distance estimation in localization scenarios makes it necessary to use velocity estimation, which could cope with these situations. Therefore, we propose to employ *time warping*, a dynamic programming scheme commonly used in machine learning and speech processing.

One of the presented methods, called *dynamic time warping* (DTW), was invented in the late 70 s initially either for aligning digitized samples of words pronounced by different speakers for recognition purposes [14] or for the alignment of biological sequences [15]. In recent years, a further development of this method, *correlation optimized warping* (COW), was proposed for the alignment of chromatographic profiles and spectra [16–18]. It was first suggested in 1998 as a way to correct chromatograms for shifts in the time axis prior to multivariate modelling and was based on incorporation of a cross-correlation correction step.

### *3.1. Dynamic Time Warping*

*Dynamic Time Warping* (DTW) is commonly used for comparing data sequences, time series or classification samples. Certain test samples are compared with a reference sample by stretching the signal by duplicating distinctive data points. As a measure of quality, a cost function, *e.g.* the sum or squared sum, is minimized in an optimal way. Its efficient implementation by means of dynamic programming makes it widely used in machine learning applications, *e.g.* optical character recognition [19] or speech recognition as well as in robotics [20] and medical applications [21].

In this paper, we propose using DTW to determine the rail vehicle velocity by finding the time shift between the two sensor heads as a distortion necessary to realign the two sample

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 161–173.

signals. To clarify the basic idea of optimality by minimizing the cost function between the signals, a short description of DTW is given below.

The distance $D$ between two signals $s_1(t)$ and $s_2(t)$ is defined by (4). As a cost function, the simple absolute distance is chosen:

$$D(s_1(t_i), s_2(t_j)) = \sum_{i=j=1}^{n} \left| (s_1(t_i) - s_2(t_j)) \right| = \sum_{i=j=1}^{n} d[i, j]. \tag{4}$$

A constant time lag or non-stationarities in frequency and phase lead to large distances although the signals could be quite similar. The DTW algorithm eliminates this difference as it enables to keep a given data sample for several steps, *i.e.* stretching the signal to be compared, until the distance between the signals is minimized. An illustrative, quantitative example is shown in Fig. 5. It shows the two signals and the resulting path chosen to minimize the cost function in a so-called distance matrix. To achieve the minimization in an optimal sense, a so-called cost path $W$ is introduced, which is defined as a series of indexed pairs:

$$\begin{aligned} W &= (w_1, w_2, \ldots w_K) \quad \text{with} \quad n \le K \le 2n-1, \\ w_k &= [i_k; j_k] \qquad\qquad 1 \le i_k; j_k \le n. \end{aligned} \tag{5}$$

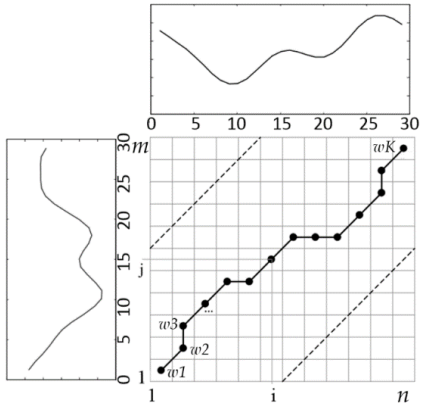



Fig. 5. A simulated cost path when warping a signal (image taken from [22]).

Thus, as shown in the Fig. 5, the individual sample points are reused to stretch the signal:

$$\begin{aligned} &s_1(t_1) \quad s_1(t_2) \quad s_1(t_3) \quad s_1(t_4)\ldots, \\ &s_2(t_1) \quad s_2(t_1) \quad s_2(t_2) \quad s_2(t_3)\ldots, \end{aligned} \tag{6}$$

which results in the corresponding path:

$$W = ([1; 1], [2; 2], [3; 2], [4; 3]\ldots). \tag{7}$$

To find the path $W_{opt}$ with the lowest costs the optimization problem is formulated as follows:

$$D = \min_W \sum_{[i_k, j_k] \in W} d[i_k, j_k] = \sum_{[i_k, j_k] \in W_{opt}} d[i_k, j_k]. \tag{8}$$

Since not all possible cost paths are useful for the purpose of velocity estimation and in order to prevent singular solutions, the following constraints are introduced:

– Boundary conditions: $w_1 = [1; 1], w_k = [n; n]$. This ensures that the start and end points of both signals are identical;

– Continuity:
given $w_{k-1} = [i_{k-1}; j_{k-1}]$ and $w_k = [i_k; j_k] \Rightarrow i_k - i_{k-1} \leq 1$ and $j_k - j_{k-1} \leq 1$. This constraint ensures that only adjacent cells can be reached in the path.
– Monotonicity:
given $w_{k-1} = [i_{k-1}; j_{k-1}]$ and $w_k = [i_k; j_k] \Rightarrow i_k - i_{k-1} \geq 0$ and $j_k - j_{k-1} \geq 0$. This last constraint forces monotonic spacing of the points regarding the time.

Direct solving the optimization problem according to (8) is not feasible because of the exponential computational complexity $O(n^n)$.

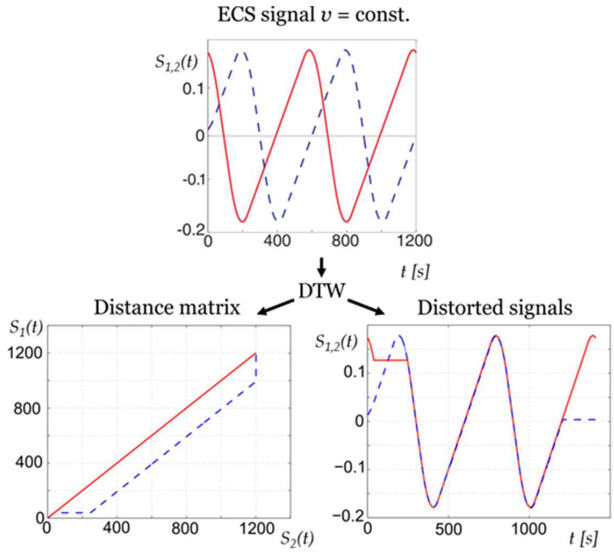



Fig. 6. The simulated result of DTW. The picture above shows ECS signals at a constant speed, where $s_1(t)$ is represented by a solid line and $s_2(t)$ by a dashed line. The corresponding distance matrix is shown below on the left, emphasizing the constant signal offset. The warped signal is shown below on the right.

Instead, a cost matrix that contains all accumulated costs along all possible paths up to the concerned cell is created. By starting at the cell with the lowest final cost, the minimal cost path is found recursively, which reduces the computational effort to $O(n^2)$. The chosen path is optimal in the sense of Bellman [4] and is closely related to the well-known Viterbi algorithm. The result of applying the algorithm on idealized ECS signals is shown in Fig. 6. One can see from the corresponding distance matrix that the algorithm aligns the signals by eliminating the constant phase shift right at the beginning.

### *3.2. Correlation Optimized Time Warping (COW)*

COW was first described as an adaptation of DTW in the field of gas chromatography [6]. In contrast to DTW, COW tries to adjust the two signals piecewise. Instead of the distance measure of (4), the signal similarity is based on a cross-correlation within the signals. To do this, the reference and target signals are divided into segments of $m$ length, each of which can be either stretched or compressed. Due to this stretching, the segments must be shifted by a certain distance $x_i$ which must satisfy the following condition:

$$x_i \pm u_i; [u_i \in (-t, t)]. \tag{9}$$

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 161–173.

After shifting the segments with a so-called slack $t$ the stretched signal is compared with the reference signal by adapting the new segment size from $m + t$ or $m - t$ to the reference signal size and this is done by means of a linear interpolation. Afterwards, the signal similarity can be determined by a cross-correlation. The possible segment shifting by the slack and the subsequent comparison of the signals is not computationally feasible even for a small amount of segments and a small shifting slack. Therefore, the problem is solved again using a recursive approach based on optimal sub-solutions. The derivation of the final algorithm is outside the scope of this contribution and is described in detail in [6] and [18].

## 4. Simulation

### *4.1. Simulation framework*

A simulation was done to verify the possibility of determining the shift of EDS signals with these warping algorithms. Therefore, several velocity profiles were simulated assuming a sleeper distance of 600 mm, a sensor distance of 208 mm and a sensor sampling rate of 1 kHz. Accelerations were restricted to a maximum of 3 m/s$^2$ which is the maximum achievable breaking power of typical rail vehicles. Afterwards, Additive White Gaussian Noise was added to simulate real-world disturbances. The sequences were chosen to have a length of 1–2 seconds which corresponds to the common correlator length. Simulated velocity profiles and their respective noise-free signals are shown in Fig. 7.

### *4.2. Simulation results*

Given the simulated signals and velocity profiles, all examined algorithms, DTW, COW and the classical cross-correlation were tested with three scenarios: noisy signals at a constant velocity, noise-free signals with accelerations and noisy signals with accelerations.

**Cross-Correlation**

The results for the cross-correlation showed the expected behaviour. Noise-free signals at a constant velocity are reliably processed and white noise does not reduce the quality considerably. The simulated result for CCF with a distinctive peak (due to finite sample lengths) is shown in Fig. 8a. The case is different for a simulated starting scenario, shown in Fig. 8b. For the simulated acceleration of 2 m/s$^2$ one would expect an end velocity of 3 m/s, yet the correlation estimates an average velocity of 1.4 m/s within the interval. A decreasing correlation interval is not a viable solution because at least two rail clamp signals are needed to have a reliable estimate.

**Classical DTW**

The results of DTW for constant velocities have already been shown in Fig. 6. The degradation in the signal quality by *Additive White Noise* is not negligible. Fig. 9 shows the results for the constant acceleration scenario. The velocity estimate quality degrades rapidly and that leads to large jumps and false values.

**COW**

The results obtained by applying the COW algorithm on simulated and noise-free signals with a constant acceleration are shown in Fig. 10. The vector lengths indicate the amount by which the segments must be shifted to be matched. They are directly proportional to the corresponding velocity. Different lengths inside an interval indicate either acceleration or braking manoeuvres.

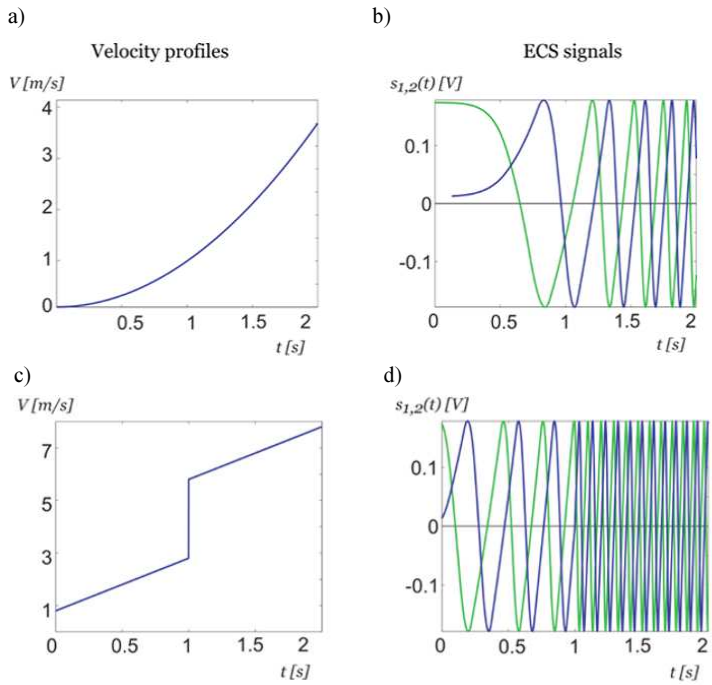



Fig. 7. The simulated ECS signals; (a) and (c) show the simulated velocity profiles; (b) and (d) show the corresponding signals of two sensor coils without additive noise.

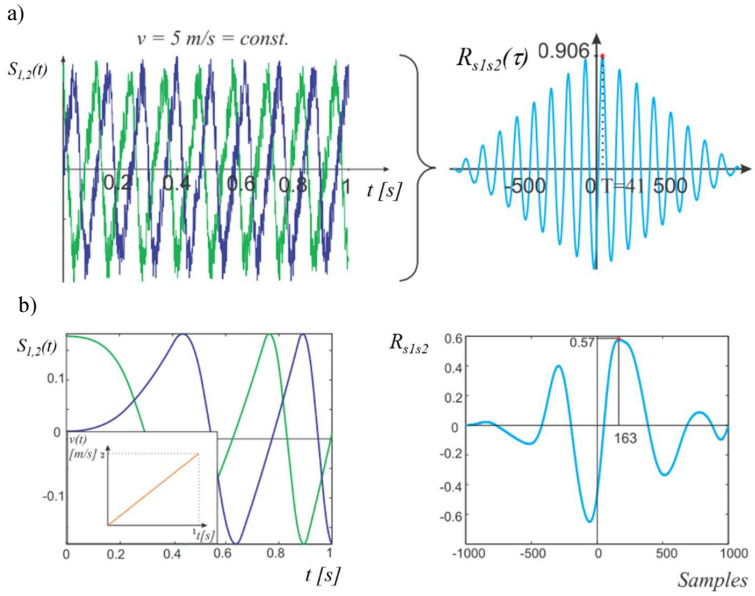



Fig. 8. The qualitative results of the cross-correlation applied to simulated data. The cross-correlation for noisy constant-velocity ECS signals (a); The cross-correlation for a linear acceleration (b).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 161–173.

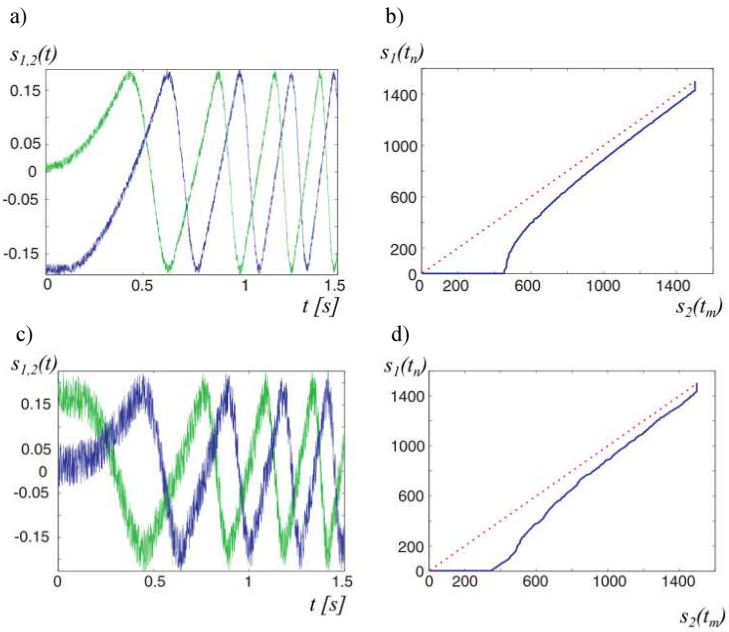



Fig. 9. The simulated results for DTW. The drawings on the left depict the input signals for a constant acceleration with increasing noise; the drawings on the right show the estimated velocity.

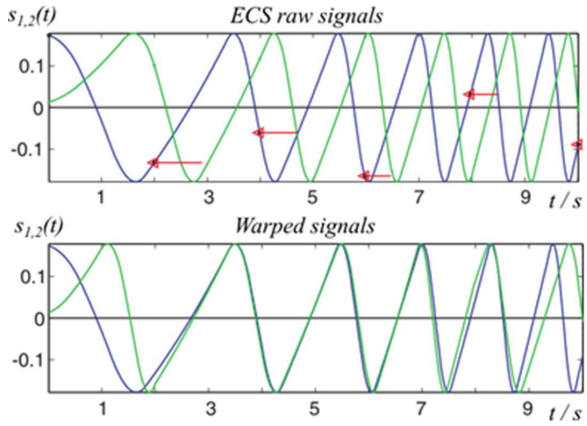



Fig. 10. Simulated result of COW. The upper section shows simulated eddy current signals with accelerations. The arrows indicate the shift of the individual segments. The lower section shows the warped results.

Adding moderate noise does not change the results at all. As the intrinsic cross-correlation quality measure COW is much more robust against additional noise than the classic DTW. Fig. 11 shows the results for noisy signals at a constant acceleration and clearly demonstrates capabilities of the approach.

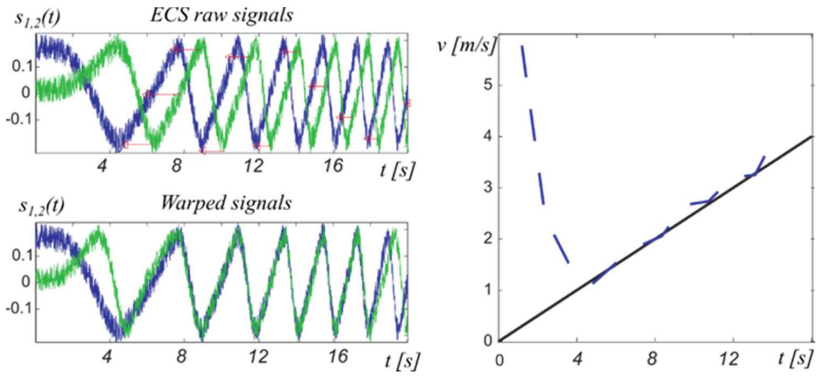



Fig. 11. The simulated results for COW given noisy signals with a constant acceleration. The left section shows the input and warped signals; the right section shows the estimated velocity in the segments as a dotted line and the correct profile as a solid line.

Summary: The results clearly indicate usefulness of the COW algorithm as compared to the DTW algorithm and the classic cross-correlation. It is robust against noise and can deal with an acceleration within the interval. As a drawback it should be mentioned that the computational load largely exceeds that of the cross-correlation.

Table 1 gives a qualitative overview of the obtained results.

Table 1. Qualitative Comparison of the simulated data.

| | Complexity of computation | Robustness against noise | Precision v = const. / ≠ const. |
|---|---|---|---|
| CLC | ++ | + | + / - |
| DTW | - | -- | + / -- |
| COW | -- | + | + / + |

## 5. Experimental results

The algorithms were also used for real-world data obtained during test drives on a tram. Fig. 12 shows an experimental signal sample and the resulting velocity estimates for CLC and DTW. The cross-correlation approach shows good estimation behaviour. The signal jumps marked in Fig. 12c correspond to abrupt velocity changes in estimation and indicate a low applicability to real-world scenarios, where a more reliable and smooth result is needed.

On the other hand, COW could solidify the expectations based on the simulated data, showing a good overall noise reduction and reliable and smooth velocity estimates even in areas with high accelerations. This is shown in Fig. 13 for a sequence with a nearly constant velocity (a) and a sequence for a starting train in a station (b). The signals align well in both scenarios and exemplify the quality of the velocity estimation. The results at constant velocities are 4.14 m/s for CLC and 4.42 m/s for COW, which principally proves applicability of the warping algorithm. To cite quantitative results for areas with higher accelerations an additional velocity sensor is necessary, as the classical CLC fails to deliver comparable true ground data.

The results obtained with the experimental data correspond mostly with the results of the simulation. An exception is the DTW performance, which shows a strong decrease in precision when compared to the simulation and other methods.

The results clearly indicate that COW is an alternative to the common CLC-based velocity estimation, especially at low velocity manoeuvers. A drawback is its high computational load.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 161–173.

Intermediate results cannot be calculated in advance and up to several seconds are needed, even for small sequences. This makes the presented algorithms less capable for online systems than the model-based approaches recently presented in [23] and [24]. Nonetheless, as proposed in our system setup in Fig. 1, the warping-based estimate needs only to be calculated below a certain velocity threshold or can be used to get precise velocity results in an offline mapping step. As a rough initial estimate is given by CLC, one can even optimize the necessary segment length to make the computation feasible. This could spread capabilities of the ECS system for additional low-speed use cases like turnout detection and classification, which were proposed in [25].

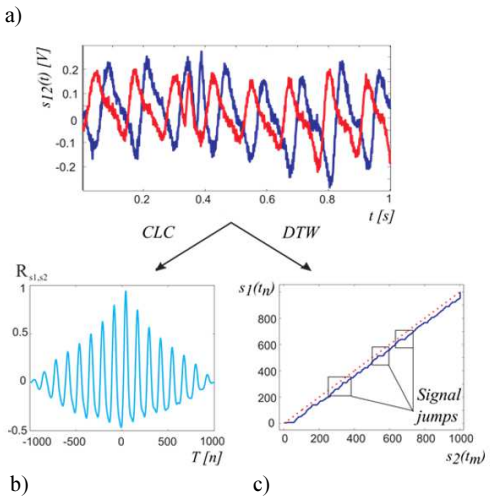



Fig. 12. CLC (b); DTW (c); results for the experimental data (a).

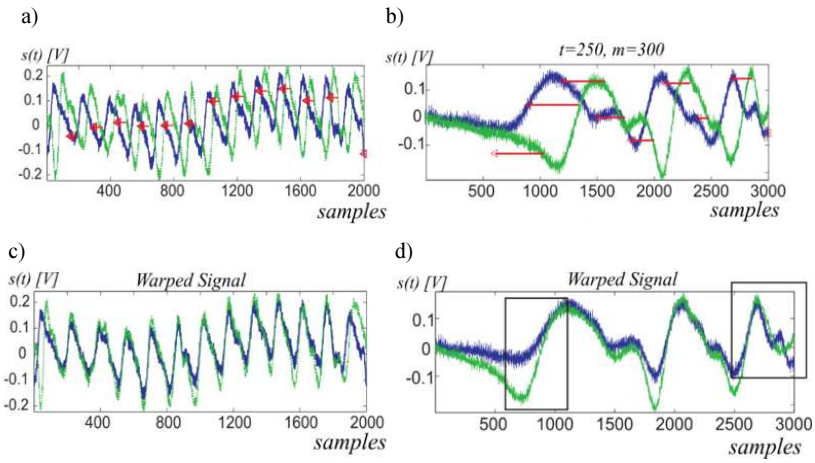



Fig. 13. The results for COW with the experimental data. (c) shows the warped signal from (a) at a nearly constant velocity. (d) represents the situation at a high acceleration for the input signal (b).

## 6. Conclusion

This paper proposes a novel approach for determining the signal shift of ECS signals for the purpose of velocity estimation. Dynamic time warping and correlation-optimized warping were described. They may be applied in an additional pre-processing step for precise rail vehicle velocity estimation in low-speed scenarios. The important conclusion is that simple signal warping should be handled with care. One must bear in mind that the classical DTW was originally proposed for pattern recognition tasks and is not robust to noise or larger signal variations. COW copes much better with the given challenges in heavy-duty train operating systems. We have qualitatively and quantitatively demonstrated that a good velocity estimate in low-speed and high-acceleration scenarios is possible.

The proposed system chooses an algorithm based on the current velocity. The fast and reliable CLC estimate is used in areas with a low acceleration, whereas the COW one is incorporated when the rail vehicle either starts or comes to a halt.

COW is not particularly sensitive to the exact choice of parameters, although maximum segment and slack lengths should not be exceeded to ensure quality. Due to a relatively small search space, the trial and error approach for choosing the optimal settings is feasible even on modest computer systems. Further work could examine the best parameters and additional speeding up of the algorithm.

## References

[1] Winter, P., Braband J., de Cicco, P. (2009). *Compendium on ERTMS: European Rail Traffic Management System*. UIC International.

[2] Böhringer, F. (2003). Train location based on fusion of satellite and trainborne sensor data. *Location Services and Navigation Technologies*, 5084, 76–85.

[3] Engelberg, T., Mesch, F. (2000). Eddy current sensor system for non-contact speed and distance measurement of rail vehicles. *Computers in Railways VII*, 1261–1270.

[4] Bellman, R.E. (1957). *Dynamic Programming*. Princeton University Press, Princeton, New Jersey.

[5] Sakoe, H., Chiba, S. (1978). *Dynamic programming algorithm optimization for spoken word recognition.* 26, 43–49.

[6] Tomasi, G., van den Berg, F., Anderson, C. (2004). Correlation optimized warping an dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemomertrics*, 18, 231−241.

[7] Moll, H., Burkhardt, H. (1979). A modified Newton-Raphson-Search for the Model-Adaptive Identifications of Delays. *Identification and System Parameter Estimation*.

[8] Sadowski, J. (2014). Velocity Measurement using the Fdoa Method in Ground-Based Radio Navigation System. *Metrol. Meas. Syst.*, 21(2), 363–376.

[9] Kowalczyk, A., Hanus, R., Szlachta, A. (2011). Investigation of the Statistical Method of Time Delay Estimation Based on Conditional Averaging of Delayed Signal. *Metrol. Meas. Syst.*, 18(2), 335–342.

[10] McIntire, P., McMaster, R.C. (1986). *Nondestructive Testing Handbook*. The American Society for Nondestructive Testing, Columbus, Ohio.

[11] Hensel, S., Hasberg, C. (2008). HMM Based Segmentation of Continuous Eddy Current Sensor Signals. *Proc. of the 11th IEEE International Conference on Intelligent Transportation Systems*, 760−765.

[12] Papoulis, A., Unnikrishna Pillai, S. (2002). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Boston.

[13] Berger, C. (1998). Optische Korrelationssensoren zur Geschwindigkeitsmessung technischer Objekte. *VDI Verlag*.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 161–173.

[14] Kruskall, J., Liberman, M. (1983). The symmetric time-warping problem: from continuous to discrete. *Time Warps, String Edits, and Mocromolecules: The Theory and Practice of Sequence Comparison*, 125–161.

[15] Needleman, S., Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, 443–453.

[16] Wang, C.P., Isenhour, T.L. (1987). Time-warping algorithm applied to chromatographic peak matching gas-chromatography Fourier-transform infrared mass-spectrometry. *Anal.Chem*., 59, 649–654.

[17] Reiner, E, Abbey, L.E., Moran, T.F., Papamichalis, P., Shafer, R.W. (1979). Characterization of normal human cells by pyrolysis gas-chromatography mass spectrometry. *Biomedical mass spectrometry (Biomed Mass Spectrom)*, 6, 491–498.

[18] Nielsen, N. Carstensen, J., Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatoraphic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, A 805(1–2), 17–35.

[19] Bahlmann, C. (2003). *Dynamic Time Warping techniques on an example of the on-line handwriting recognition*. Department of Computer Science, Albert-Ludwigs-University Freiburg.

[20] Schmill, M., Oates, T., Cohen, P. (1999). Learned models for continuous planning. *Seventh International Workshop on Artificial Intelligence and Statistics*.

[21] Caiani, E. *et al.* (1998). Warped-average template technique to track on a cycle-by-cycle basis the cardiac filling phases on left ventricular volume. *IEEE Computers in Cardiology*.

[22] Keogh, E., Pazzani, M. (2000). Derivative Dynamic Time Warping. *Technical Report, University of California.*

[23] Strauss, T., Hasberg, C., Hensel, S. (2009). Correlation based velocity estimation during acceleration phases with application in rail vehicles. *IEEE/SP 15th Workshop on Statistical Signal Processing*.

[24] Hensel, S., Strauss, T., Marinov, M. (2015). *Eddy current sensor based velocity and distance estimation in rail vehicles.* IET Science, Measurement & Technology, 9(7), 875–888.

[25] Hensel, S., Hasberg, C., Stiller, C. (2011). Probabilistic Rail Vehicle Localization with Eddy Current Sensors in Topological Maps. *IEEE Transactions on Intelligent Transportation Systems*, 12(4), 1–13.

www.czasopisma.pan.pl PAN www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 175–184.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# SEMI-AUTOMATIC APPARATUS FOR MEASURING WETTING PROPERTIES AT HIGH TEMPERATURES

**Marcin Bąkała[1)], Rafał Wojciechowski[1)], Dominik Sankowski[1)], Adam Rylski[2)]**

*1) Lodz University of Technology, Institute of Applied Computer Science, Stefanowskiego 18/22, 90-924 Łódź, Poland (✉ m.bakala@kis.p.lodz.pl, +48 42 631 2750, r.wojciechowski@kis.p.lodz.pl, d.sankowski@kis.p.lodz.pl)*

*2) Lodz University of Technology, Institute of Materials Science and Engineering, Stefanowskiego 1/15, 90-924 Łódź, Poland (adam.rylski@p.lodz.pl)*

**Abstract**

Determination of the physico-chemical interactions between liquid and solid substances is a key technological factor in many industrial processes in metallurgy, electronics or the aviation industry, where technological processes are based on soldering/brazing technologies. Understanding of the bonding process, reactions between materials and their dynamics enables to make research on new materials and joining technologies, as well as to optimise and compare the existing ones. The paper focuses on a wetting force measurement method and its practical implementation in a laboratory stand – an integrated platform for automatic wetting force measurement at high temperatures. As an example of using the laboratory stand, an analysis of Ag addition to Cu-based brazes, including measurement of the wetting force and the wetting angle, is presented.

Keywords: surface tension, wetting angle, wetting force, measurement system.



## 1. Introduction

Research on new materials and joining technologies is one of the most important areas of interest in materials' engineering. Good knowledge of the structure of materials, their physical and chemical properties, behaviour in various conditions enables to design new materials or technologies as well as to customize the existing ones. In the latter case, they become more economical and offer a better quality. This research is also fostered by European Union directives (Restriction of Hazardous Substances 2011/65/EU; Waste Electrical and Electronic Equipment 2012/19/UE) determining the detailed requirements of use of certain substances in the electrical and electronic devices, which are in operation in many countries. It also forces changes in industrial technologies [1, 2]. Current research on the modern materials and their industrial applications involves development of new measurement methods providing accurate, quantitative information on the behaviour of a material in various technological processes. Modification of a material structure has to meet the requirements regarding specified properties of new materials in appropriate usage conditions. Frequently, the difficulty lies in application of a new technology in the production process, where elements made of modern materials are assembled with those of traditional ones.

Determination of the physico-chemical interactions between liquid and solid substances is a key technological factor in many industrial processes in metallurgy, electronics or the aviation industry. The primary phenomenon used in the bonding process is wetting the joined surfaces with a liquid metal, which is described by the primary interfacial impact parameters, *i.e.* the wetting force and the wetting angle [3–7]. Information on the values and dynamics of the above-mentioned parameters can be obtained by performing experiments based on the immersion method. Knowledge of the wetting dynamics enables to customize the existing

Article history: received on Mar. 17, 2016; accepted on Aug. 28, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0004.

technologies, optimize them or fit to appropriate process conditions and requirements (significant reduction of the process time and temperature, use of a protection atmosphere, application of fluxes, surface preparation, and the like). In addition, it enables to study materials and joining technologies based on soldering.

## 2. General concept of measurement and analysis of wetting force

A wetting force measurement procedure, consisting in observation of a specimen's weight changes during an experiment of immersing a specimen in a liquid braze, is known as the Wilhelmy plate method. A high-precision scales system measures the resultant forces acting on the vertical specimen. Based on analysis of the distribution of the forces acting on a specimen before and after the immersion (Fig. 1), the capillary wetting force is denoted as [3–7]:

$$F_c = F_{m2} - F_{m1} + S_a \rho g h \tag{1}$$

and the wetting angle:

$$\theta = \arccos\left(\frac{F_c}{P_a \sigma_{LV}}\right), \tag{2}$$

where: $F_c$ – the capillary wetting force; $F_{m1}$, $F_{m2}$ – the forces registered by the scales system before and after immersion, respectively; $S_a$ – the cross-sectional area; $\rho$ – the solder density; $g$ – the gravity acceleration; $h$ – the immersion depth; $P_a$ – the specimen perimeter; $\sigma_{LV}$ – the solder surface tension.

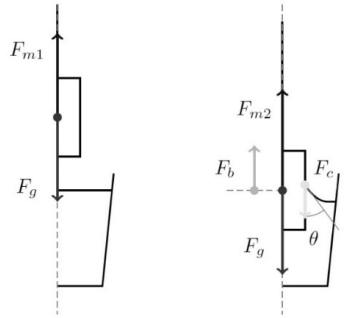



Fig. 1. Distribution of the forces acting on a specimen before and after its immersion in a fluid braze.

Equation (2) applied to a specific liquid-solid system describes a single wetting angle. In fact, there are many metastable interstates on the solid-liquid boundary arising from the material heterogeneity, impurities and surface structure modification, which affect differences of the wetting angle relative to the aforementioned angle value. When the phase boundary is moving, instead of static contact angles, one should measure dynamic contact angles, and thus determine the range of angles referred to the advancing and receding angles, which creates the wetting angle hysteresis [8].

## 3. Integrated platform for automatic wetting force measurement at high temperatures

### *3.1. Computer system overview*

The integrated platform for automatic wetting force measurement at high temperatures is an autonomous stand enabling complex research on the dynamic brazing process properties –

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 175–184.

the wetting force at temperatures of up to 1000°C with various technological gas atmospheres. An overview of the measurement stand and its most important subsystems – heating, driver and loading systems, is presented in Fig. 2 [9].

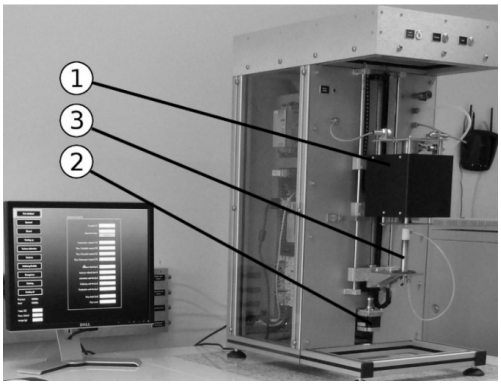



Fig. 2. An overview of the wetting force measurement system (1 – heating; 2 – driver; 3 – loading subsystems).

The system offers a wide range of working temperatures up to 1000°C with a braze bath temperature control accuracy of 0.5°C and a precisely controlled gas protective/reductive atmosphere (usually based on argon, nitrogen and hydrogen of 5N purity). The measurement cycle is fully automated, the process parameters and experiment sequence are controlled by a real-time automated system. The measurement procedure is based on analysis of the forces acting on a specimen during its immersion in a liquid braze. The integrated platform is a unique solution for industrial and laboratory purposes enabling to make research on design of new materials and soldering/brazing technologies or to optimize and verify the existing ones [9]. The commercial solutions available on the market (*e.g.* Metronelec Menisco products) offer wetting force measurements in a gas-protective atmosphere mostly at a temperature of only up to 450°C [10]. The measurement system presented in this paper enables to examine the dynamic brazing process properties at up to 1000°C.

### *3.2. System architecture*

The architecture of the integrated platform for automatic wetting force measurement at high temperatures consists of functional blocks responsible for executing separated tasks. The basic subsystems include: heating, gas, scales and loading/drive subsystems, all supervised by an industrial software *programmable logic controller* (PLC) equipped with extension cards necessary to communicate with industrial parts/devices. An additional component of the research stand is an autonomous information system enabling to analyse the experiment results. The architecture of the research stand is presented in Fig. 3 [9].

The main part of the measurement system is a custom-built heating system based on a cylindrical resistance furnace (Fig. 4, A) enabling to heat a specimen up to 1000°C with a temperature control accuracy of 0.5°C. The heating system is controlled by a PLC software coupled with an appropriate power driver and a thermocouple receiving temperature from the furnace chamber. During the experiment, the pot with a braze material is placed in the centre of the furnace, on an appropriate base (Fig. 4, B). Location of the pot in the furnace and movement of the whole heating system in the direction of the fixed specimen are executed by a custom-built driver system (Fig. 4, G) controlled also by the PLC controller. It ensures

a position control accuracy of 0.05 mm and a speed of up to 200 mm $s^{-1}$. The specimen is detected at the furnace entrance by the optical detection system (Fig. 3, D) using Pepperl+Fuchs photoelectric sensors [11]. The contact between the specimen and the liquid braze is recognized by rapid specimen weight changes when the specimen face touches the solder surface. The balance subsystem is based on a Mettler Toledo weighing module with the maximum load of 220 g and accuracy of 0.1 mg [12]. During the heating some additional phenomena, *e.g.* oxidation, can occur. Therefore, the measurement system is equipped with a gas-protective atmosphere based on nitrogen (Fig. 4, E). A hydrogen and argon mixture is applied as a reduction atmosphere in the furnace chamber. The gas flow is controlled by the Brooks *mass flow controller* (MFC) elements calibrated for specified external gas sources with the maximum flow of 250 mln $min^{-1}$ and control accuracy below 1% of the actual flow (20–100% of full scale) [13]. All components, including sensors, motors and controllers, are managed by a WAGO software PLC based on an x86 family processor, Linux operating system and CodeSys real-time environment. The PLC unit is equipped with extension cards including digital/nalog inputs/outputs, a stepper motor controller, RS interfaces and others [14].

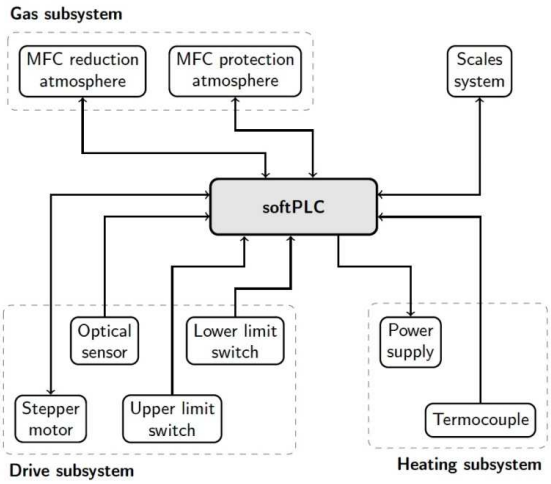



Fig. 3. The architecture of the research stand with reference to heating, gas, drive, scales subsystems and controlling devices.

### *3.3. Flow diagram of wettability measurement experiment*

The sequence of tasks executed by all hardware system components according to the process parameters set by the user, which present the required values of the main parameters of the process, *i.e.* the temperature, the gas flows, *etc*., is illustrated in Fig. 5. Selected process parameters that can be set by the user are presented in Table 1. The experiment tasks refer to the appropriate states of the device and define all actions and conditions necessary to proceed to the next experiment step. The experiments can differ in details; however, their general process outline is similar. During the experiment, the current process variables (*i.e*. weight, temperature, specimen position, time) are recorded.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 175–184.

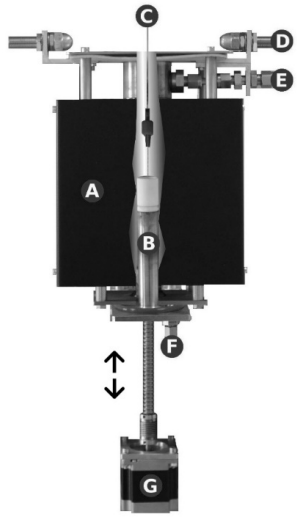



Fig. 4. A cross-section of the measurement system (A – the heating system; B – the base and the pot with a liquid braze; C – the tested specimen and its mounting; D – the specimen detection system; E, F – the gas supply system; G – the driver system).

Table 1. Selected process parameters that can be set by the user.

| Parameter | Description |
|---|---|
| Temperature [°C] | Process temperature (up to 1000 °C) |
| Immersion depth [mm] | Specimen immersion depth in fluid solder / braze |
| Immersion speed [mm $s^{-1}$] | Specimen immersion speed |
| Emergence speed [mm $s^{-1}$] | Specimen removal speed |
| Step depth [mm] | Specimen immersion step depth in fluid solder / braze (multiple immersion steps' experiment) |
| Step count | Specimen immersion step count (multiple immersion steps' experiment) |
| Activation time [s] | Time of specimen stay in the furnace interior, over a solder / braze bath, before specimen immersion |
| Stabilization time [s] | Time of specimen stay in fluid solder / braze |
| Cooling time [s] | Time of specimen stay in the furnace interior, over a solder / braze bath, after specimen removal |

The experiment is carried out at a high temperature, in the presence of a protective gas atmosphere. Stabilization of the thermal conditions and gas reduction atmosphere before the main part of the experiment is vital. The experiment procedure starts with a vertical movement of the furnace with the pot filled with a braze towards the specimen – until the specimen is detected at the furnace entrance by the optical sensor (Fig. 7, pos. A). After the detection, the movement still continues, until the specimen is placed in the furnace chamber at a fixed position to activate the specimen surface for a required time. Next, the furnace vertical movement towards the specimen continues until the contact with the braze surface is detected (Fig. 7, pos. B). The contact is detected by the scales system – when the face of the specimen reaches the braze bath, rapid weight changes are observed. After the contact detection, the specimen is placed at an appropriate depth, where it stays for a required stabilization time (Fig. 7, pos. D–F). Finally, after the specimen is taken out of the braze and kept above its surface for

an appropriate time, the specimen is removed from the furnace, at which point the experiment ends [9].

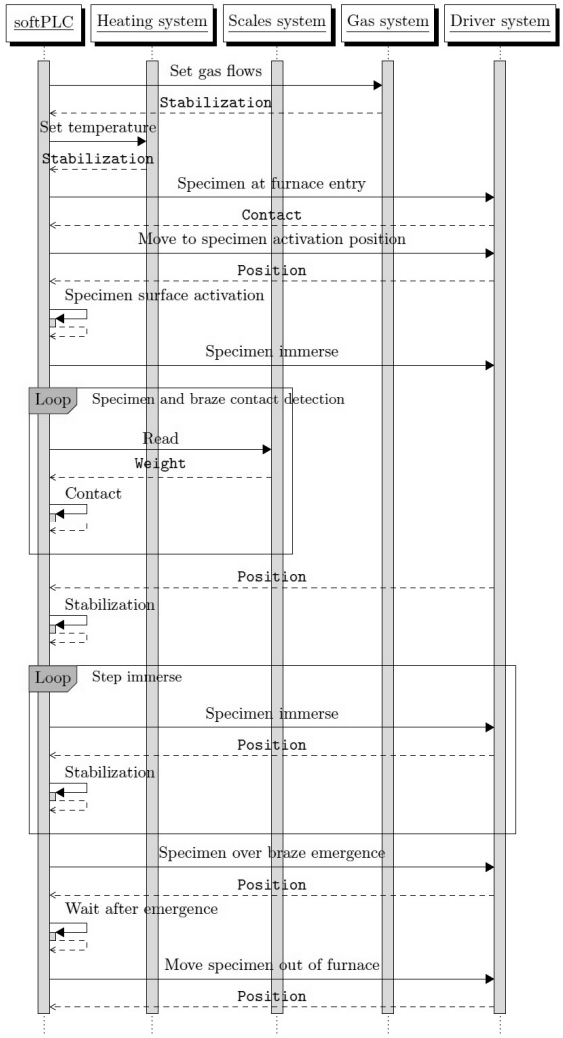



Fig. 5. A flow diagram of the measurement experiment.

The apparatus, due to a wide range of process parameters that can be set by the user, offers a great flexibility in research planning. The proper selection of parameters affects the correctness and purpose of executed experiment. The most important parameters determining the type of study are timing parameters (immersion and emergence speed, stabilization time) enabling to focus on measurement of the wetting dynamics (*e.g.* advancing and receding angles) or observation of the wetting changes during the specimen's static stay in a fluid braze. The correct determination of most available parameters requires sufficient experience from the operator, and it should be preceded with execution of a series of testing experiments. For example, the time of the specimen's stay in a fluid braze should match, in most cases, the stabilization of the wetting force. Some of the parameters, even those fundamental to the

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 175–184.

process, such as temperature or gas flows, can be verified after the experiment with an additional study, *e.g.* microstructure research on oxides or on the existence of specific phases. A number of technical parameters are not available to the user, because they are related to the activity and protection of the research stand (component devices reading frequency, speed and position limits, *etc.*), but they should be taken into account when planning the experiment flow and other parameters.

## 4. Example of analysis of Ag addition to Cu-based brazes

As an example of using the integrated platform for automatic wetting force measurement at high temperatures, an analysis of Ag addition to Cu-based brazes was made. The immersion experiments for the Cu specimens and Cu-based brazes with Ag addition, commonly used in electrotechnical and mechanical applications, were carried out for three different brazes (braze 1–93.8% Cu, 6.2% P; braze 2–91.8% Cu, 6.2% P, 2% Ag; braze 3–80% Cu, 5% P, 15% Ag). Parameters of the experiments are shown in Table 2. The wettability changes during the immersion process of copper specimens in liquid brazes are presented in Fig. 6.

Table 2. Parameters of the immersion experiments.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Temperatures [°C] | 740–760 | Activation time [s] | 20 |
| Immersion depth [mm] | 5 | Stabilization time [s] | 5 |
| Immersion speed [mm $s^{-1}$] | 5 | Cooling time [s] | 10 |

A detailed analysis of the wetting force changes during the specimen's static stay in a fluid braze is presented in Fig. 7. The main stage of the experiment starts with contacting the specimen front surface with the liquid braze (Fig. 7, pos. B). During the specimen sinking process, the acting buoyancy force rises linearly with an immersion depth. At the same time, the braze bath surface is deflected and a down-curved meniscus is formed. The wetting angle $\theta$ is changing until it reaches its maximum value and forms an obtuse angle (Fig. 7, pos. C). The temperature of the specimen rises to the braze temperature. Bonds between the two-phase atoms and the SL-phase boundary are created. The C–F stages correspond to the wetting progress. The capillary force is increasing, with its value equal to the buoyancy force at point D and with a value equal to 0 at point E, where the wetting angle reaches 90°. From point E, the wetting angle forms an acute angle towards the metastable equilibrium value $\theta^0$, where the resultant force acting on the specimen remains in balance. At point F, the wetting force reaches 90% of its maximum value, which is needed to create a joint with good properties. The last stage is the emergence process (Fig. 7, pos. G). The liquid meniscus is broken off and the experiment is ended (Fig. 7, pos. H). The dynamic parameters of the tested braze materials are shown in Table 3. Additionally, research on the surface tension at a required temperature was performed for each braze material using the lying drop method implemented in the ThermoWet device [15]. Knowledge of the surface tension parameter is needed to calculate the wetting angle changes as a function of time (2). The wetting angle changes registered for Cu-based brazes with Ag addition are presented in Fig. 8.

The immersion experiments were carried out for Cu specimens and three Cu-based brazes specified above. The experiment results prove that addition of Ag in the composition of brazes affects the wetting dynamics and parameters of braze materials and process conditions. The addition of Ag in a braze composition causes reduction of the wetting force – for braze 1 the registered wetting force is about 7.84 mN and the metastable equilibrium wetting angle is about 58°, for braze 3 the force value is 4.67 mN and the required wetting angle 72°. For brazes with

a lower amount of Ag addition, the wetting dynamics is better – for braze 1 and braze 2 the calculated dynamics ratio is about 0.3, but for braze 3 it exceeds 0.4. The dynamics ratio is a basic parameter for technology design which specifies the speed of the wetting process (a lower value of the ratio denotes a rectangular shape of the wetting curve for rapid wetting, while a value close to 1 indicates a slow increase in the wetting force). However, increasing the Ag component lowers the process temperature (the recommended operating temperature for braze 1 is 760°C, but for braze 3–730°C) [16].

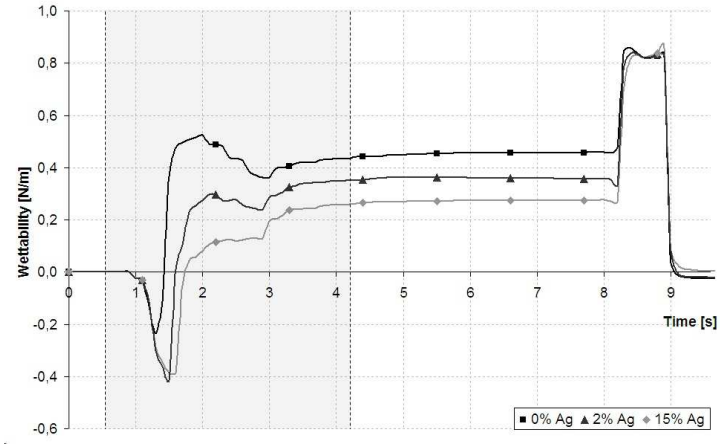



Fig. 6. The wettability changes registered for Cu-based brazes with Ag addition described above (braze 1 – ■; braze 2 – ▲; braze 3 – ♦) with an area selected for a detailed analysis of dynamics.

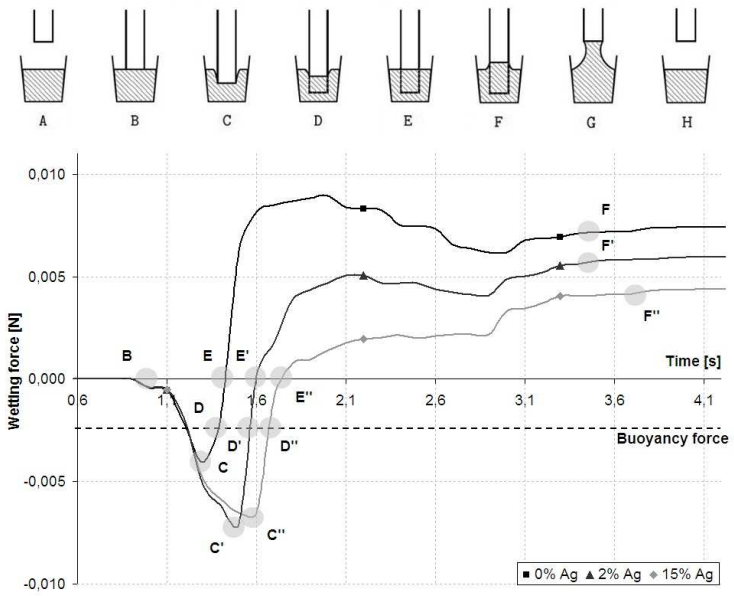



Fig. 7. The wetting force changes and specific points for Cu-based brazes with Ag addition (braze 1 – ■; braze 2 – ▲; braze 3 – ♦) within the marked area (Fig. 6).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 175–184.

Table 3. The wetting dynamics parameters of the immersion experiments for Cu flat specimens and Cu-based brazes with Ag addition.

| Parameter | Braze 1 | Braze 2 | Braze 3 |
|---|---|---|---|
| $t_0$ [s], point B | 1.0 | 1.0 | 1.0 |
| $t_{\theta max}$, point C | 1.3 | 1.4 | 1.6 |
| $t_{FB}$, point D | 1.4 | 1.5 | 1.7 |
| $t_{\theta = 90°}$, point E | 1.5 | 1.6 | 1.8 |
| $t_{90\%}$ [s], point F | 3.4 | 3.3 | 3.8 |
| $t_C$ [s] | 7.8 | 7.8 | 7.8 |
| $F_G$ [mN] | 7.84 | 6.0 | 4.67 |
| $(t_{90\%} - t_0) / (t_C - t_0)$ | 0.35 | 0.34 | 0.41 |

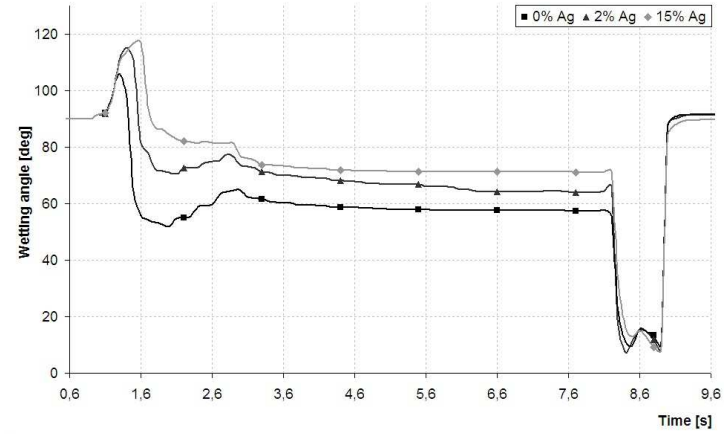



Fig. 8. The wetting angle changes registered for Cu-based brazes with Ag addition (braze 1 – ■; braze 2 – ▲; braze 3 – ♦).

## 5. Conclusions

Determination of conditions for the production of good quality material joints is nowadays carried out by means of equilibrium wettability tests. In the paper, a methodology of wetting force measurement, including a general concept of the immersion experiment and calculation of its parameters, is presented. The methodology was used in the automatic research stand – an integrated platform for automatic determination of high-temperature braze wettability, enabling to perform a comprehensive study of dynamic properties of brazes at temperatures of up to 1000°C with the use of various technological gas atmospheres. The research carried out on an integrated platform supplemented with an additional analysis of the microstructure provides comprehensive information about material properties and their behaviour that can be used for industrial and laboratory purposes regarding the design of new joining technologies and materials, optimization of brazing and soldering process parameters as well as verification of quality of the existing technologies.

In the paper, an example of research on Ag addition to Cu-based brazes is presented. The immersion experiments carried out for Cu specimens and Cu-based brazes (braze 1–93.8% Cu, 6.2% P; braze 2–91.8% Cu, 6.2% P, 2% Ag; braze 3–80% Cu, 5% P, 15% Ag) prove that addition of Ag in the composition of brazes affects the wetting dynamics and parameters

of braze materials and process conditions. The addition of Ag in a braze composition causes a reduction of the wetting force and a slowdown of the wetting dynamics. On the other hand, increasing the Ag component lowers the process temperature.

## References

[1] Directive 2011/65/EU of the European Parliament and of the Council of 8th June 2011 on the restriction of the use of certain hazardous substances in electrical and electronic equipment (RoHS2). *Official Journal of the European Union*. L 174 / 88.

[2] Directive 2012/19/UE of the European Parliament and of the Council of 4th July 2012 on waste electrical and electronic equipment (WEEE2). *Official Journal of the European Union*. L 197/38.

[3] Eustanthopoulos, N., Nicholas, M.G., Drevet, B. (1999). *Wettability at High Temperatures*. Pergamon Press.

[4] Humpston, G., Jacobson, D.M. (1993). *Principles of Soldering and Brazing*. ASM.

[5] Bormashenko, E. (2013). *Wetting of Real Surfaces*. Studies in Mathematical Physics. De Gruyter.

[6] Rivollet, I., Chatain, D., Eustanthopoulos, N. (1990). Simultaneous measurement of contact angles and work of adhesion in metal-ceramic systems by the immersion-emersion technique. *J. of Material Science*, 25, 3179–3185.

[7] Adamson, A.W. (1997). *Physical Chemistry of Surfaces*. John Wiley & Sons.

[8] Yuan, Y., Lee, T.R. (2013). Contact Angle and Wetting Properties. *Surface Science Techniques*, 51, 3–34.

[9] Integrated platform for automatic wetting force measurement at high temperatures. Information about the laboratory stand. *http://www.kis.p.lodz.pl*. (Feb. 2016).

[10] Metronelec Menisco ST88. *http://www.metronelec.biz.* (Jun. 2016).

[11] Pepperl+Fuchs. Manufacturer of photoelectric sensors. *http://www.pepperl-fuchs.com*. (Jun. 2016).

[12] Mettler Toledo. Manufacturer of weighing modules. *http://www.mt.com.* (Jun. 2016).

[13] Brooks. Manufacturer of mass flow meters and mass flow controllers. *http://www.brooksinstrument.com.* (Jun. 2016).

[14] WAGO. Manufacturer of PLC and automation modules. *http://www.wago.com.* (Jun. 2016).

[15] ThermoWet. Information about the laboratory stand. *http://www.kis.p.lodz.pl.* (Jun. 2016).

[16] Massalski, T.B. (1986). *Binary alloys phase diagrams*. ASM.

www.czasopisma.pan.pl PAN www.journals.pan.pl
POLSKA AKADEMIA NAUK

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 185–192.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# NEW TYPE OF PIEZORESISTIVE PRESSURE SENSORS FOR ENVIRONMENTS WITH RAPIDLY CHANGING TEMPERATURE

**Myroslav Tykhan[1)], Orest Ivakhiv[1)], Vasyl Teslyuk[2)]**

*1) Lviv Polytechnic National University, Institute of Computer Technologies, Automation and Metrology, Kniazia Romana 19, Lviv,79013, Ukraine (✉ tykhanm@ukr.net, +38 067 704 5082, oresti@polynet.lviv.ua)*

*2) Lviv Polytechnic National University, Institute of Computer Science and Information Technology, Mytropolyta Andreja 5, Lviv,79013, Ukraine (tesliuk@polynet.lviv.ua)*

**Abstract**

The theoretical aspects of a new type of piezo-resistive pressure sensors for environments with rapidly changing temperatures are presented. The idea is that the sensor has two identical diaphragms which have different coefficients of linear thermal expansion. Therefore, when measuring pressure in environments with variable temperature, the diaphragms will have different deflection. This difference can be used to make appropriate correction of the sensor output signal and, thus, to increase accuracy of measurement. Since physical principles of sensors operation enable fast correction of the output signal, the sensor can be used in environments with rapidly changing temperature, which is its essential advantage. The paper presents practical implementation of the proposed theoretical aspects and the results of testing the developed sensor.

Keywords: pressure, sensor, non-stationary, temperature.



## 1. Introduction

Nowadays numerous technical systems need pressure sensors with high accuracy of measurement in environments with rapidly changing non-stationary temperature, such as aerospace systems, scientific research, *etc*. [1–4]. Therefore, development of more advanced sensors and methods of accurate pressure measurement with fast correction of the temperature error is a very important task.

Achievements in microelectronic technology have brought about a large group of piezo-resistive sensors designed for different environments [5–9]. Analysis of the characteristics of such sensors suggests their high accuracy, because the temperature error of some types of sensors is equal to fractions of a per cent [7–11]. However, when measuring pressure in environments with rapidly changing temperature (thermal shock, *etc*.), the error can exceed 30% [12, 13].

The current methods [14–18] (such as thermal compensation, cooling method, application of thermal protection films, combined measurement of temperature and pressure, *etc*.) can reduce temperature error. However, when temperature is changing rapidly, the error is significant [12, 13]. In addition, fast correction of errors, which is required for automatic control systems, cannot be performed.

## 2. Theoretical aspects

Some theoretical aspects for round plates are discussed below [19]. The vertical deflection in the centre of a thin round rigidly restrained plate (Fig. 1) loaded with a pressure $p$ is:

Article history: received on Feb. 12, 2016; accepted on Sep. 29, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0010.

$$w = \frac{pR^4 12(1-\nu^2)}{64Eh^3} = p\frac{1.66}{\gamma\omega^2}, \tag{1}$$

where: $R$ is a radius of the plate; $h$ is a thickness of the plate; $E$ is the elasticity modulus; $\nu$ is the Poisson's ratio; $\omega$ is a natural frequency of the round plate; $\gamma = \rho h$; $\rho$ is a density of the plate's material.

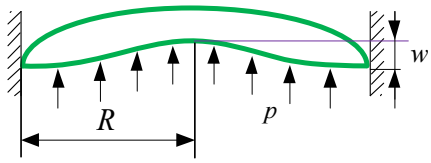



Fig. 1. The vertical deflection of a round plate.

If the plate is loaded with a pressure $p$ and radial forces of compression/tension $\pm N_r$ (Fig. 2), then the vertical deflection in the plate's center is equal to:

$$w = p\frac{1.66}{\gamma\omega^2 \pm N_r \dfrac{10.24}{R^2}}. \tag{2}$$

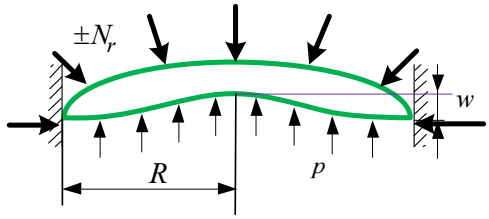



Fig. 2. The vertical deflection of a round plate under radial forces.

Let us assume that we have two thin round restrained plates 1 and 2 (Fig. 3) with identical geometrical parameters and physical characteristics of the materials, except for the coefficient of linear thermal expansion ( $\lambda_1 < \lambda_2$ ).

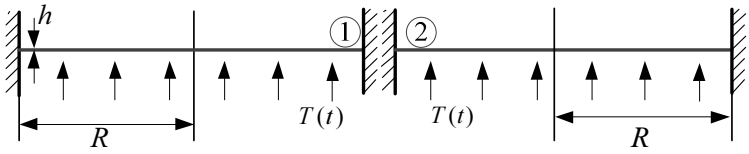



Fig. 3. Identical round plates with different coefficients of linear thermal expansion of materials.

If plates 1 and 2 (Fig. 3) are under influence of a non-stationary temperature, the thermal stress that will arise in them will be expressed as:

$$\sigma_{r1}(z,t) = \sigma_{\varphi 2}(z,t) = \frac{-E\lambda_1}{1-\nu} T(z,t) \tag{3}$$

and

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 185–192.

$$\sigma_{r2}(z,t) = \sigma_{\varphi 2}(z,t) = \frac{-E\lambda_2}{1-\nu} T(z,t)\,, \tag{4}$$

where $T(z, t)$ is a temperature field in the body of the plates with their coefficients of linear thermal expansion $\lambda_1$ and $\lambda_2$.

In the case of restrained plates 1 and 2, the above-mentioned thermal stresses (3) and (4) correspond to the radial forces $\pm N_r$ (Fig. 4) as follows:

$$N_{r1}(t) = \int_0^h \sigma_{r1}(z,t)dz = \frac{\pm E\lambda_1}{1-\nu} \int_0^h T(z,t)dz \tag{5}$$

and

$$N_{r2}(t) = \int_0^h \sigma_{r2}(z,t)dz = \frac{\pm E\lambda_2}{1-\nu} \int_0^h T(z,t)dz\,. \tag{6}$$

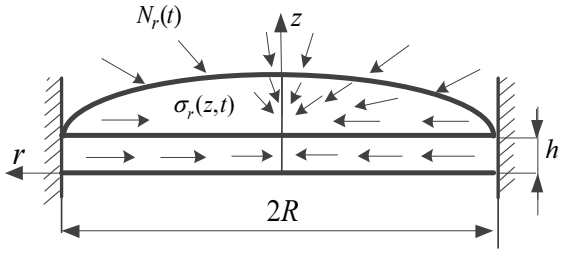



Fig. 4. The thermal stresses and corresponding radial forces.

It is apparent that, as the geometrical parameters and physical properties of the materials of plates 1 and 2 (Fig. 3) are identical, their static and dynamic characteristics in standard conditions will be identical, too. However, under a thermal impact, as it is seen from the obtained (2) – (6), due to the different coefficients of linear thermal expansion, the plates will be exposed to different thermal stresses, which is why their statics and dynamics at measurement of a pressure $p(t)$ will differ.

If plates 1 and 2 are under a pressure $p$ and radial forces $N_{r1}(t)$ and $N_{r2}(t)$, then the vertical deflections in the centre of the plates are:

$$w_1(t) = p_0 \frac{1.66}{\gamma\omega^2 \pm N_{r1}(t)\dfrac{10.24}{R^2}} \tag{7}$$

and

$$w_2(t) = p_0 \frac{1.66}{\gamma\omega^2 \pm N_{r2}(t)\dfrac{10.24}{R^2}}\,. \tag{8}$$

Taking into account (5) and (6) will result in the following equations:

$$w_1(t) = p_0 \frac{1.66}{\gamma\omega^2 \pm \dfrac{10.24}{R^2}\dfrac{E\lambda_1}{1-\nu}\displaystyle\int_0^h T(z,t)dz}\,, \tag{9}$$

$$w_2(t) = p_0 \frac{1.66}{\gamma\omega^2 \pm \frac{10.24}{R^2}\frac{E\lambda_2}{1-\nu}\int_0^h T(z,t)dz}. \tag{10}$$

Solving (9) and (10), we will obtain:

$$p_0(t) = \frac{\gamma\omega^2}{1{,}66} w_1(t)\left[1 \pm \frac{\lambda_1\left(w_2(t) - w_1(t)\right)}{w_1(t)\lambda_1 - w_2(t)\lambda_2}\right]. \tag{11}$$

These are the above-described regularities that underlay the proposed type of pressure sensors and the corresponding method of pressure measurements.

## 3. Practical implementation of theoretical aspects

Let us use the mentioned plates as diaphragms in a piezo-resistive pressure sensor (Fig. 5). On diaphragms 1 and 2, the identical piezo-resistors 3 and 4 will be set, which at a measured pressure $p(t)$ will produce output signals $U_1(t)$ and $U_2(t)$, respectively.

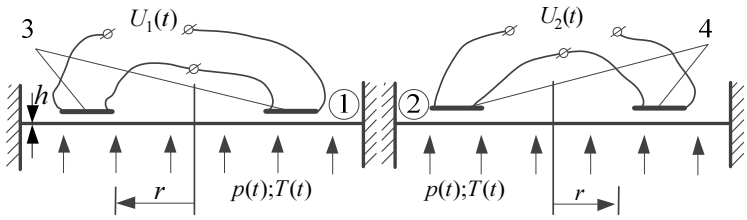



Fig. 5. A piezo-resistive pressure sensor with two diaphragms.

If there is no thermal influence, the output signals from piezo-resistors 3 and 4 will be equal to ($U_1(t) = U_2(t)$), and if the measured pressure $p(t)$ is not changing fast, it can be written that:

$$U_1(t) = U_2(t) = \frac{p(t)}{k}, \tag{12}$$

where $k$ is a static coefficient of sensor transformation, which takes into account the topology of piezo-resistors on diaphragms and their supply voltage.

Thus, in this case the values of the output signals $U_1(t)$ or $U_2(t)$ make it possible to define the value of measured pressure as:

$$p(t) = U_1(t) \times k = U_2(t) \times k\,. \tag{13}$$

If the diaphragms during measurement of the pressure $p(t)$ are under influence of a temperature $T(t)$ (Fig. 5), the obtained values $U_1(t)$ and $U_2(t)$ will differ, since the diaphragms, having different values of coefficients of linear thermal expansion, will be exposed to different thermal stresses. Therefore, the values of measured pressure $p_1(t)$ and $p_2(t)$ determined by the output signals $U_1(t)$ or $U_2(t)$ will also be different:

$$p_1(t) = U_1(t) \times k \neq p_2(t) = U_2(t) \times k\,. \tag{14}$$

On the other hand, using expression (1), we will have:

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 185–192.

$$p_1(t) = \frac{w_1(t)\gamma\omega^2}{1.66}, \tag{15}$$

$$p_2(t) = \frac{w_2(t)\gamma\omega^2}{1.66}. \tag{16}$$

Taking into account (14–16), (11) will be written down as follows:

$$p_0(t) = k \cdot U_1(t)\left[1 \pm \frac{\lambda_1\left(U_2(t) - U_1(t)\right)}{U_1(t)\lambda_1 - U_2(t)\lambda_2}\right]. \tag{17}$$

This equation enables to obtain the true value of measured pressure in an environment with a non-stationary temperature.

If the measured pressure changes quickly, then the output signal of the sensor will also have a dynamic error. Therefore, the first step in the measurement method must be elimination of this error.

The dynamic model of piezo-resistive pressure sensors with a round diaphragm is known to be described by integral Volterra equation:

$$U(t) = k\int_0^t p(\tau)\cdot e^{-\beta(t-\tau)}\cdot \sin(\omega(t-\tau))d\tau\,. \tag{18}$$

Double integration of (18) will result in:

$$p(t) = \frac{\ddot{U}(t) + 2\cdot\beta\cdot\dot{U}(t) + (\omega^2 + \beta^2)\cdot U(t)}{k\cdot\omega}. \tag{19}$$

Direct implementation of this equation is impossible, since direct differentiation of the output signal is an incorrect procedure. Therefore, application of the known methods of measuring dynamic pressure [20, 21] will result in $p_1(t)$ and $p_2(t)$ for each of the diaphragms. Then, to find the value of measured pressure, (17) will be used, in which the just defined values $p_1(t)$ and $p_2(t)$ are substituted for the values $U_1(t)$ and $U_2(t)$.

The proposed type of piezo-resistive pressure sensor with two diaphragms for environments with a non-stationary temperature is shown in Fig. 6.

The piezo-resistive pressure sensor for environments with a non-stationary temperature (Fig. 6b) consists of a body 4, round diaphragms 1 and 2 rigidly restrained in the body 4, piezo-resistors 3 and 5 located on diaphragms 1 and 2.

The measurement goes through the following stages:

- the piezo-resistive pressure sensor with two diaphragms with identical parameters but with different coefficients of linear thermal expansion perceives the measured pressure $p_0(t)$;
- the output signals $U_1(t)$ and $U_2(t)$ of the sensor are processed using any of the known methods of dynamic pressure measurement [6,7] in order to correct the dynamic error, and the values $p_1(t)$ and $p_2(t)$ are obtained;
- the true value of measured pressure $p_0(t)$ is calculated by (17), in which the defined values $p_1(t)$ and $p_2(t)$ are substituted for $U_1(t)$ and $U_2(t)$, respectively;
- if the pressure is not dynamic, then the values of output signals $U_1(t)$ and $U_2(t)$ of the sensor are directly placed in (17), and the true value of measured pressure $p_0(t)$ is calculated.

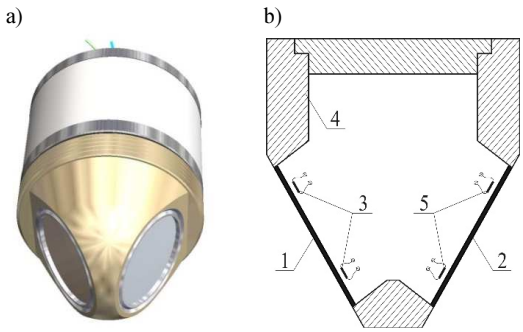



Fig. 6. A piezo-resistive pressure sensor with two diaphragms for environments with a non-stationary temperature (a) and its design (b).

A piezo-resistive pressure sensor for environments with a non-stationary temperature (Fig. 6b) consists of a body 4, round diaphragms 1 and 2 rigidly restrained in the body 4, piezo-resistors 3 and 5 located on diaphragms 1 and 2.

## 4. Testing the sensor and the corresponding method

The method and the respective sensor were tested in the conditions of simultaneous application of a pressure shock with an amplitude of 0.2 MPa and a thermal shock with an amplitude of 135°C.

The sensor (Fig. 6) had two diaphragms with the radius $R$ = 4 mm, thickness $h$ = 0.21 mm and elasticity modulus $E = 2.1 \cdot 10^{11}\ Pa$, the diaphragms were made of an alloy with different coefficients of linear thermal expansion: $\lambda_1 < \lambda_2$ and the Poisson's ratio $\nu = 0.3$.

When measuring a pressure shock with simultaneous applying a thermal shock, the output signals $\tilde{U}_1(t)$ and $\tilde{U}_2(t)$ are obtained (Fig. 7).

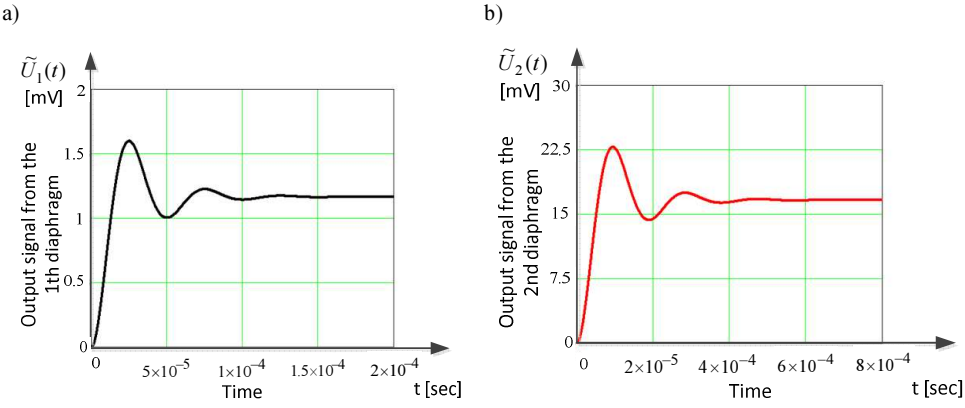



Fig. 7. The output signals of the sensor at measurement under a pressure shock with simultaneous applying a thermal shock: the diaphragm of an alloy with $\lambda_1$ (a); the diaphragm of an alloy with $\lambda_2$ (b).

Using the dynamic pressure measurement method [20] results in the restored signals $\tilde{p}_1(t)$ and $\tilde{p}_2(t)$ (Fig. 8).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 185–192.

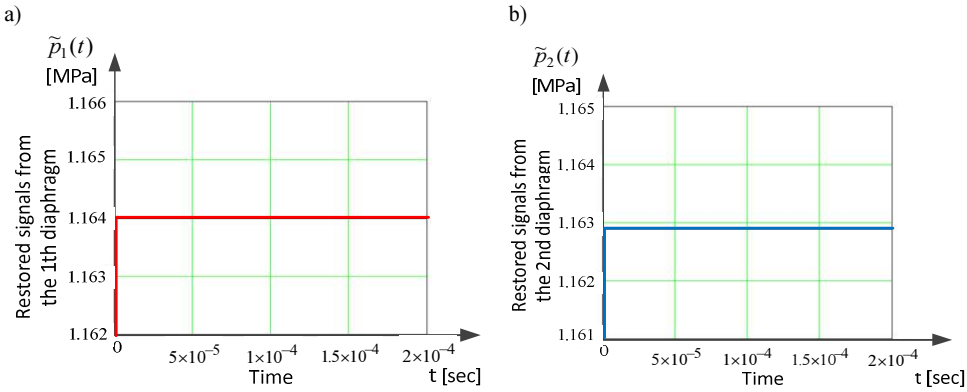



Fig. 8. The restored signals $\tilde{p}_1(t)$ and $\tilde{p}_2(t)$: the diaphragm of an alloy with $\lambda_1$ (a); the diaphragm of an alloy with $\lambda_2$ (b).

Using (17), we obtained the true value of measured pressure (Fig. 9), and the relative error equalled 2,45%.

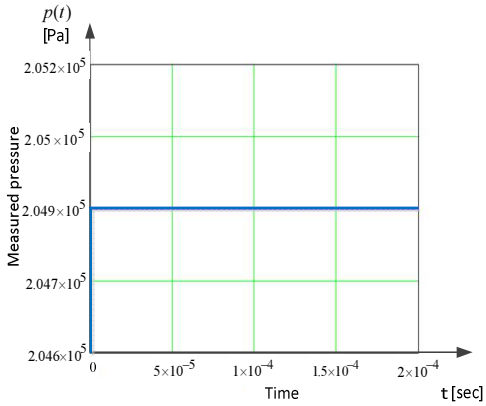



Fig. 9. The true value of measured pressure.

## 5. Conclusions

Therefore, testing the method and the corresponding sensor gave a quite satisfactory result.

The formula (17) enables to calculate the true values of both dynamic and static pressures of an environment with a non-stationary temperature. Since such a calculation is implemented in the real-time mode due to the simplicity of mathematical procedures, accurate measurement of pressure can be carried out at a rapidly changing ambient temperature. Thus, the proposed type of piezo-resistive pressure sensor with two diaphragms can be used in high-speed control systems. However, a weak point of such sensors is the necessity to ensure identical properties and parameters of the diaphragms.

## References

[1] *Mems for Automotive and Aerospace Applications* (2013). Woodhead Publishing Series in Electronic and Optical Materials №32. Edited by M. Kraft and Neil M. White. Woodhead Publishing Limited.

[2] Markelov, I.G. (2009). Complex of pressures sensors for exploitation on the objects of atomic energy. *Sensors and systems*, 11/12, 24–25.

[3] Custom Pressure Sensors for the Aerospace Industry. Merit Sensor. *https://meritsensor.com.*

[4] Sensors for Aerospace & Defense. PCB Piezotronics. *https://www.pcb.com/aerospace.*

[5] Zhang, J., Liu, Q., Zhong, Y. (2008). A Tire Pressure Monitoring System Based on Wireless Sensor Networks Technology. *International Conference on MultiMedia and Information Technology*, 602–605.

[6] Lai, C.C., Dai, C.L., Chang, P.Z. (2005). A Piezoresistive Micro Pressure Sensor Fabricated by Commercial DPDM CMOS Process. *Tamkang Journal of Science and Engineering*, 8(1), 67–73.

[7] Kistler. Measure, analyze, innovate. *https://www.kistler.com.*

[8] Piezoresistive_vs_Piezoelectric. *https://www.kulite.com.*

[9] *https://www.omega.com.*

[10] Schatz, O. (2004). Recent trends in automotive sensors. *Proc. of IEEE Sensors*, 1, 236–239.

[11] Ko, H.S. (2007). Novel fabrication of a pressure sensor with polymer material and evaluation of its performance. *Journal of Micromechanics and Microengineering*, 17(8), 1640–1648.

[12] Mokrov, E.A., Belozubov, E.M., Tikhomirov, D.V. (2004). Minimizing the error of thin-film piezoresistive pressure sensors under the influence of non-stationary temperature. *Sensors and systems*, 1, 26–29.

[13] Mokrov, E.A., Vasilev, V.A., Belozubov, E.M. (2005). Application thermoprotective films for minimizing the influence of non-stationary temperature on thin-film piezoresistive pressure sensors. *Sensors and systems*, 9, 21–23.

[14] Kasten, K., Amelung, J., Mokwa, W. (2000). CMOS-compatible capacitive high temperature pressure sensors. *Sensors and Actuators A: Physical*, 85, 147–152.

[15] Pressure transducers and Melt Pressure Sensors. *https://www.dynisco.com.*

[16] Gridchin, A., Antonov, A.A. (2012). Termocompensation of tensoresistive sensors using bipolar junction transistor in extended temperature range. Conference Paper. *Actual Problems of Electronics Instrument Engineering (APEIE), 11th International Conference*.

[17] Peng, K.H., Uang, C.M. (2003). The temperature compensation of the silicon piezo-resistive pressure sensor using the half-bridge technique. *Proc. of SPIE – The International Society for Optical Engineering*, 534.

[18] Reverter, F., Horak, A., Bilas, G., Forner, V.G., Gasulla, M. (2009). Novel and low-cost temperature compensation technique for piezoresistive pressure sensors. *XIX IMEKO World Congress. Fundamental and Applied Metrology*, Lisboa, 2084–2087.

[19] Timoshenko, S.P., Woinowsky-Krieger, S. (1979). *Theory of Plates and Shells*. New York: McGraw-Hill.

[20] Tykhan, M. (2009). *Method of measuring of dynamic pressure*. Patent of Ukraine. No. 45961, Bul. No. 23.

[21] Tykhan, M. (2009). *Method of measuring of dynamic pressure*. Patent of Ukraine. No. 88936, Bul. No. 23.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 193–199.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# PARTIAL SHADING DETECTION IN SOLAR SYSTEM USING SINGLE SHORT PULSE OF LOAD

**Mateusz Bartczak**

*Wroclaw University of Technology, Chair of Electronic and Photonic Metrology, Boleslawa Prusa 53/55, 50-317 Wroclaw, Poland*
*(✉ mateusz.bartczak@pwr.edu.pl,* +48 513 390 725*)*

**Abstract**

A single photovoltaic panel under uniform illumination has only one global maximum power point, but the same panel in irregularly illuminated conditions can have more maxima on its power-voltage curve. The irregularly illuminated conditions in most cases are results of partial shading. In the work a single short pulse of load is used to extract information about partial shading. This information can be useful and can help to make some improvements in existing MPPT algorithms. In the paper the intrinsic capacitance of a photovoltaic system is used to retrieve occurrence of partial shading.

Keywords: maximum power point, partial shading, load pulse, test station, I–V curve.



## 1. Introduction

A growing demand for energy, especially electricity, ensuring security of supply and reducing undesired effects of climate-related emissions of carbon dioxide and other pollutants resulting from combustion of fossil fuels into the atmosphere is one of the biggest economic and environmental challenges in the world in recent years. Today, the solar energy production is one of the fastest growing industries in the world and one of the fastest growing energy technologies [1]. All these processes stimulate the development of new techniques.

One of the major challenges in *photovoltaic* (PV) systems is increasing the amount of produced energy. The shape of I–V curve depends on many factors, like temperature, spectral irradiance, total irradiance, orientation to the light source, type of bypass, parasitic impedances (due to a length of cable bundles between a test station and modules), mechanical aspects, optical alignment effects. Partial shading is a serious problem and leads to lowering the efficiency of produced energy. There are several papers that describe how large is the impact of this phenomenon on the amount of produced energy [2, 3]. The authors of [4] present a MATLAB-based modelling and simulation scheme suitable for studying the I–V and P–V characteristics of a PV array under a non-uniform insolation due to partial shading. In most cases, from a hardware point of view, partial shading is handled by bypass diodes, however a method described in [5] slightly increases the efficiency of photovoltaic module by replacing bypass diodes with FET transistors of a low drain-source resistance. This solution reduces energy losses caused by the current flowing over the bypass. From a software point of view, we can find a lot of direct and indirect algorithms searching the *maximum power point* (MPP) [6]. Most of them are inappropriate to track MPP in partially shaded conditions, because the photovoltaic array characteristic curves exhibit multiple local maxima. In the literature, the most popular is a *perturb and observe* (P&O) algorithm. This approach is often used because of its simple implementation. On the other hand, a drawback of this method is its low efficiency in fast changing insolation conditions. The authors of [9, 10] expand a standard solar

Article history: received on Nov. 08, 2016; accepted on Jan. 14, 2017; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0016.

system with installing additional temperature and light sensors. They also propose a hybrid of P&O and open-voltage algorithms. Those sensors improve the system immunity for conditions of partial shading in comparison with the reference P&O method. One of the drawbacks of this solution is an additional wiring necessary to measure the temperature and irradiance.

The paper concentrates on a method dedicated to detection of partial shading conditions. The obtained results can be useful for improvement of the MPPT algorithms. The proposed method is based on analysis of the system response to a short load pulse. This work was inspired by [5, 6], where short pulses of light are used to acquire the I–V curve, and also by a load-pulse method used in power plants to measure the system impedance.

In the literature, a drawback caused by the intrinsic parasitic capacitance is generally neglected, but we can find MPPT algorithms which are strictly based on this property of a PV cell. For example, the authors of [7] developed a method similar to the incremental conductance [11] procedure, which additionally includes the parasitic capacitance. Fig. 1 shows a response of a solar system to a single short pulse of load; in this case it was a short current pulse. The curves from Fig. 1 have been plotted for the measurement setup described later in this paper. As can be concluded from Fig. 1, the smaller insolation, the higher the capacitance of PV system.

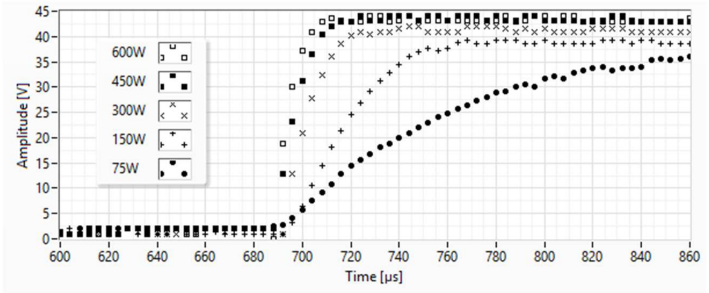



Fig. 1. A response to a short load pulse for different insolation conditions.

The rise of capacitance caused by a lower insolation can be used for detection of partial shading in a system of PV cells. Fig. 2 shows a system of two STP010-12/Kb (SUNTECH) solar panels connected in series (Table 1 gives its electrical characteristics). These panels have no additional bypass diodes but consist of 36 cells connected in series. Both solar panels are bypassed by diodes. When both panels operate in the same environmental conditions, *i.e.* the same insolation, the system response looks like one plot from Fig. 1. When insolation conditions are different for each panel then the output of both panels is the sum of their individual responses. Fig. 3 shows individual responses of each panel. As can be seen from the chart, the PV panel which is partially shaded needs more time to restore its operating voltage. Fig. 4 compares two plots – the first one is calculated as the sum of responses, whereas the second one is the response measured for both panels.

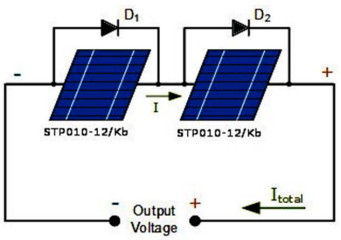



Fig. 2. Two STP010-12/Kb panels connected in series and bypass diodes.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 193–199.

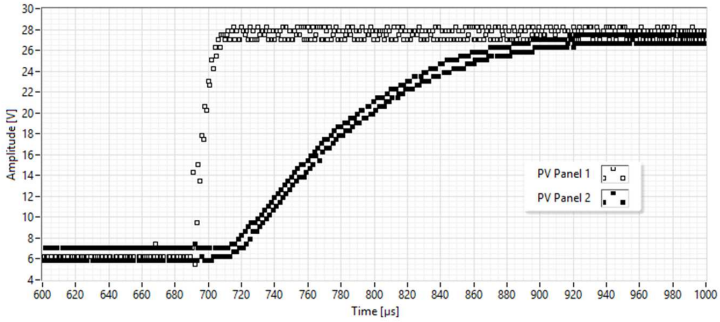



Fig. 3. The PV Panel 1 response for full insolation and the PV Panel 2 response for partly shaded conditions.

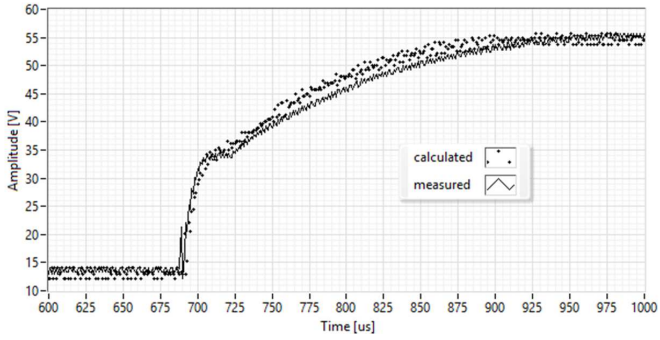



Fig. 4. The calculated and measured responses in partial shading.

Table 1. Parameters of solar panels used in the experiment.

| **Model Number** | **STP010–12/Kb** |
|---|---|
| Rated Maximum Power ($P_{MAX}$) | 10 W |
| Output Tolerance | ± 10 % |
| Current at $P_{max}$ ($I_{MP}$) | 0.57 A |
| Voltage at $P_{max}$ ($V_{MP}$) | 17.4 V |
| Short-Circuit Current ($I_{OC}$) | 0.65 A |
| Open-Circuit Voltage ($V_{OC}$) | 21.6 V |
| Nominal Operating Cell Temp. ($T_{NOCT}$) | 45ºC ± 2ºC |

## 2. Experimental setup

Figure 5 shows a simplified overview of the measurement workstation which has been used during the experiment. The setup consists of a PC with LabVIEW application, a load controller, a source of light and an oscilloscope. The LabVIEW application communicates with the load controller and the oscilloscope (Rigol D1054Z) via USB connection. The PC application, using the load controller, can set the operating voltage of solar panel and monitor the current. The test station is designed to measure the I–V curve and also to observe the response of PV system to a short load pulse . Short pulses of load are generated independently in a similar way as in the Short-Current Pulse-Base MPPT Method [8], but one change has been introduced in this setup, *i.e.* the possibility of regulating the current within the range from short-circuit to zero. Fig. 6 shows an example of the I–V curve and the power measured by the setup. The LabVIEW

software has been chosen for this experiment because it provides means for data acquisition (signal I/O), analysis and presentation, and it also supports communication between the measurement modules (*e.g.*: the oscilloscope, power supply, *etc.*). In the circuit, the load controller plays a similar role to the active load, which is based on an N-CHANNEL MOSFET transistor.

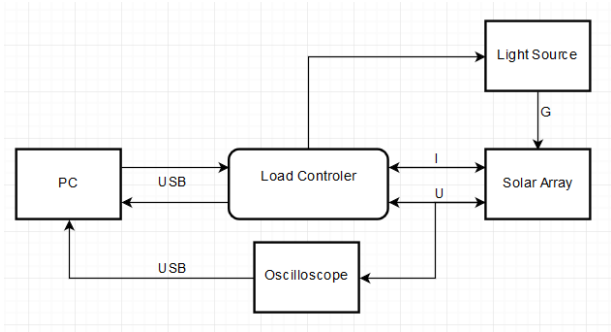



Fig. 5. A simplified scheme of measurement workstation used in the experiment.

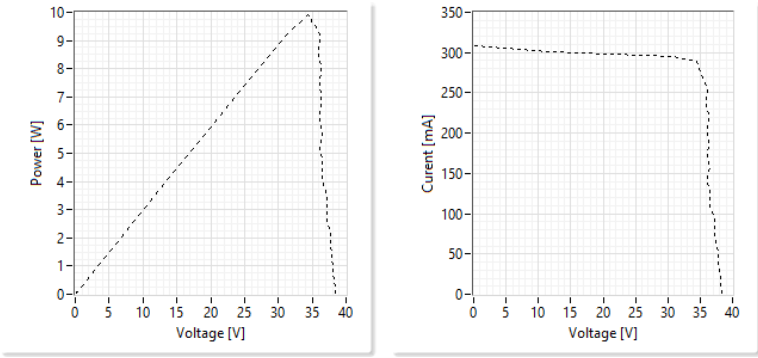



Fig. 6. Examples of curves measured in the PV system using the experimental setup: I–V curve (left), power curve (right).

## 3. Measurement results

As mentioned before, the measurements have been performed with the system (Fig. 2) consisting of two arrays of 36 solar cells connected in series. The module parameters are shown in Table 1. Every module is bypassed by a diode. Solar cells used in the experiment have a low capacitance, therefore the time required by the solar system to restore the operating voltage (when the load pulse is ended) is short. Fig. 7 shows a group of responses measured in various conditions of shading. One of the solar panels was fully exposed to irradiation, whereas the second one was gradually darkened, starting from the full exposure and ending in the complete blackout.

Successive Figs. 8–11 show the progressive darkening of one of the solar panels. The time required by a solar panel to restore its operating voltage is small for panels used in the experiment and is approximately equal to between 10 μs and 100 μs, depending on insolation conditions. The recovery time is counted starting from the moment of ending the load pulse to the moment of reaching a voltage level of $V_{oc}$. The higher capacitance of the solar system, the longer the recovery time. Fig. 12 shows the response of a 500 W system, which consists

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 193–199.

of two 250 W panels , *i.e.* a 2 × 250 W system. Thus, the time required to restore the operating voltage is approximately equal to 30 ms. When a level of insolation of two solar panels starts to be significantly inhomogeneous, the response shape starts to bend (Fig. 10 and Fig. 11). The fact of forming a bend in the response shape can be used for detecting a condition of partial shading. Fig. 13 shows plots of derivatives calculated for data from Figs. 8–11. A derivative has been approximated according to (1):

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h}. \tag{1}$$

The results for conditions of significant insolation inhomogeneity between PV panels are depicted in plots enclosed in the frame in Fig. 13. The details of their shapes can be analysed by searching data plots for peaks and valleys.

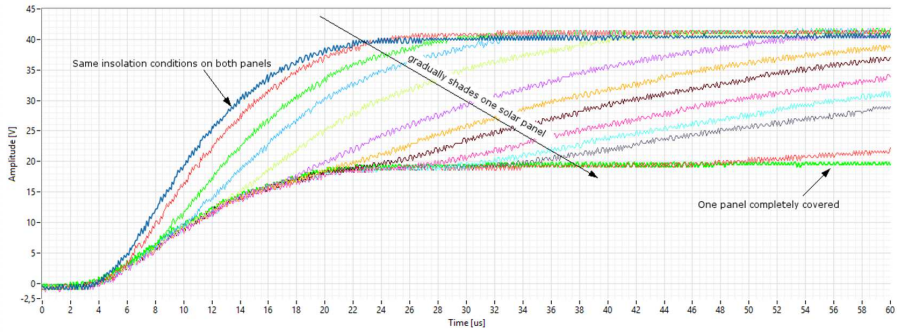



Fig. 7. A group of responses in various insolation conditions.

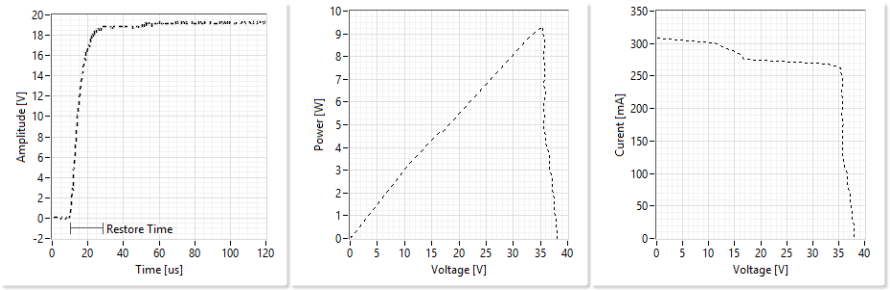



Fig. 8. The response to a short load pulse: voltage (left), power(middle), current – according to the measured I–V curve for partial shading conditions (right).

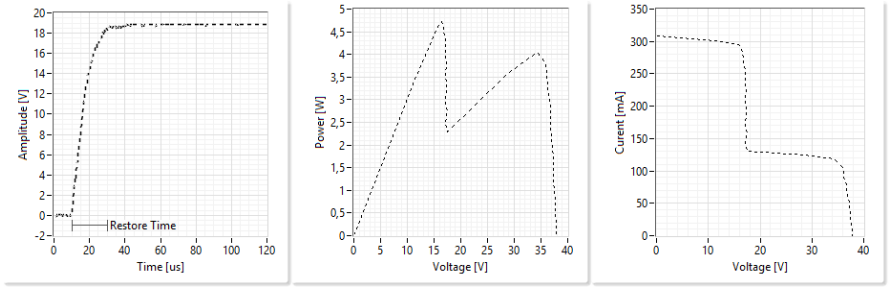



Fig. 9. The response to a short load pulse: voltage (left), power(middle), current – according to the measured I–V curve for partial shading conditions (right).

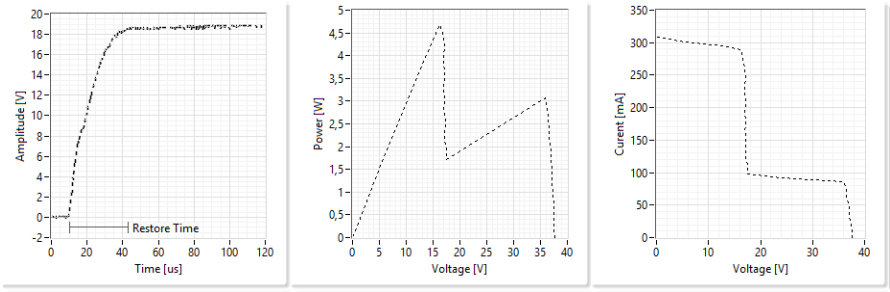



Fig. 10. The response to a short load pulse: voltage (left), power(middle), current – according to the measured I–V curve for partial shading conditions (right).

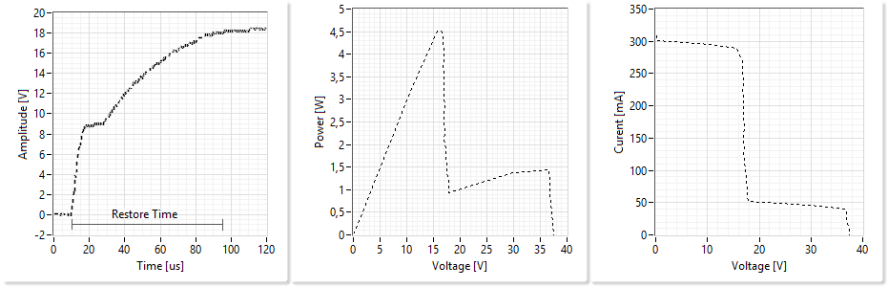



Fig. 11. The response to a short load pulse: voltage (left), power(middle), current – according to the measured I–V curve for partial shading conditions (right).

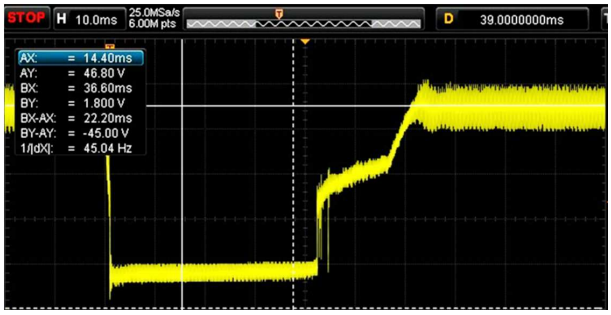



Fig. 12. The response to a short load pulse for a system consisting of larger solar panels.

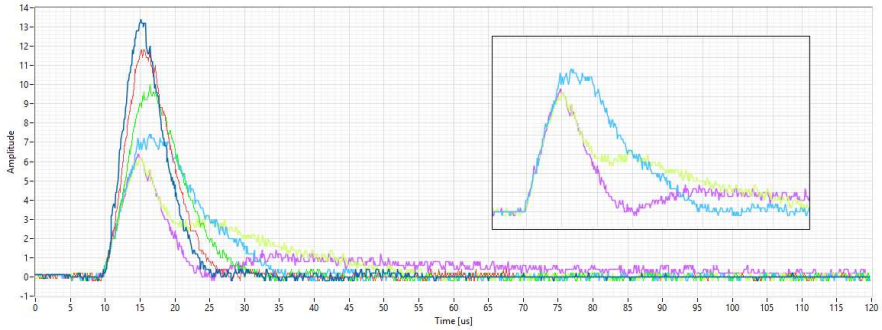



Fig. 13. Derivatives calculated from Fig. 8–11. In the highlighted frame the last three plots are obtained. for significant disproportion in insolation between PV panels.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 193–199.

## 4. Conclusion

The paper reports on a short-load-pulse method dedicated to determination of partial shading conditions in PV power generation systems. The proposed approach identifies and exploits the relationship between partial shading and the response to a short load pulse. This mechanism can be used for improving selected MPPT algorithms or constructing a hybrid of existing approaches, *e.g.* P&O and Cuckoo Search (CS) algorithms. The P&O is inefficient in conditions of partial shading, whereas the Cuckoo Search exhibits better results in partial shading conditions, but requires more time for computations. Both of them can be combined, contributing to improved power generation in conditions of partial shading. Implementation of the methodology proposed in the paper does not require sophisticated electronic circuitry, so the measurement can be performed relatively easily. On the other hand, this approach shows also a drawback, because system interruptions have to be applied quite frequently.

## References

[1] Auswertungstabellen zur Energiebilanz Deutschland 1990 bis 2014 (2016). *AG Energiebilanzen (AGEB), http://www.ag-energiebilanzen.de/.*

[2] Masoum, A.S., Padovan, F., Masoum, M.A.M. (2010). Impact of partial shading on voltage-and current-based maximum power point tracking of solar modules. *IEEE PES General Meeting*, 1–5.

[3] Balasubramanian, I.R., Ganesan, S.I., Chilakapati, N. (2014). Impact of partial shading on the output power of PV systems under partial shading conditions. *IET Power Electronics*, 7(3), 657–666.

[4] Hiren, P., Agarwal, V. (2008). MATLAB-based modeling to study the effects of partial shading on PV array characteristics. *IEEE transactions on energy conversion*, 23(1), 302–310.

[5] d'Alessandro, V., Pierluigi, G., Santolo, S. (2014). A simple bipolar transistor-based bypass approach for photovoltaic modules. *IEEE Journal of Photovoltaics*, 4(1), 405–413.

[6] Kojima, H., *et al*. (2014). Accurate and rapid measurement of high-capacitance PV cells and modules using a single short pulse light. *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*. IEEE, 2014. *Survey of Maximum PPT techniques of PV Systems*, Nasr, A.A.A., Saied, M.H., Mostafa, M.Z., Abdel- Moneim, T.M.

[7] Brambilla, A., *et al*. (1999). New approach to photovoltaic arrays maximum power point tracking. (1999). *Power Electronics Specialists Conference, PESC 99, 30th Annual IEEE*, 2.

[8] Noguchi, T., Shigenori, T., Nakamoto, R. (2002). Short-current pulse-based maximum-power-point tracking method for multiple photovoltaic-and-converter module system. *IEEE Transactions on Industrial Electronics*, 49(1), 217–223.

[9] Mroczka, J., Ostrowski, M. (2015). Maximum power point search method for photovoltaic panels which uses a light sensor in the conditions of real shading and temperature. *SPIE Optical Metrology. International Society for Optics and Photonics*.

[10] Mroczka, J., Ostrowski, M. (2014). A hybrid maximum power point search method using temperature measurements in partial shading conditions. *Metrol. Meas. Syst*., 21(4), 733–740.

[11] Safari, A., Mekhilef, S. (2011). Simulation and hardware implementation of incremental conductance MPPT with direct control method using cuk converter. *IEEE Transactions on Industrial Electronics*, 58(4), 1154–1161.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

*www.metrology.pg.gda.pl*

M&MS

# A NEW APPROACH TO SPINDLE RADIAL ERROR EVALUATION USING A MACHINE VISION SYSTEM

**C. Kavitha, S. Denis Ashok**

*VIT University, School of Mechanical Engineering, Vellore-632014, Tamil Nadu, India*
*(kavitha.c@vit.ac.in, ✉ denisashok@vit.ac.in, +91 94 4486 8585)*

**Abstract**

The spindle rotational accuracy is one of the important issues in a machine tool which affects the surface topography and dimensional accuracy of a workpiece. This paper presents a machine-vision-based approach to radial error measurement of a lathe spindle using a CMOS camera and a PC-based image processing system. In the present work, a precisely machined cylindrical master is mounted on the spindle as a datum surface and variations of its position are captured using the camera for evaluating runout of the spindle. The *Circular Hough Transform* (CHT) is used to detect variations of the centre position of the master cylinder during spindle rotation at subpixel level from a sequence of images. Radial error values of the spindle are evaluated using the Fourier series analysis of the centre position of the master cylinder calculated with the least squares curve fitting technique. The experiments have been carried out on a lathe at different operating speeds and the spindle radial error estimation results are presented. The proposed method provides a simpler approach to on-machine estimation of the spindle radial error in machine tools.

Keywords: machine vision, circular hough transform, Fourier series, runout, spindle radial error.



## 1. Introduction

A spindle is one of the key functional elements in a typical machine tool which provides a rotation to a work piece or tool. The rotational accuracy of the spindle is an important issue in production of accurate and precise components. Runout of the spindle is caused due to an installation error resulting in a misalignment of its rotational axis with either a tool or workpiece. It leads to varying a chip load on the cutting tool and machining inaccuracies related to tool positioning, causing the surface location error [1]. In early years, spindle runout tests were performed for assessing the spindle accuracy by installing a master in the spindle and measuring the *total indicated runout* (TIR) using a mechanical displacement indicator. However, the total indicated runout is not the true indicator of spindle accuracy as it is the superposition of the form error of the measured surface and the error of the spindle motion. Capacitive-sensor-based measurement techniques have been widely applied to meet the high accuracy requirement of metrology applications. A capacitive-sensor-based surface parameter evaluation method and its application to the surface finish measurement system is presented in [2]. The accuracy of spindle error measurement using capacitive sensors is affected by inherent error sources, such as a sensor offset, a thermal drift of spindle, the centring error, and the form error of the target surface installed in the spindle [3]. These methods require a measurement setup consisting of multiple numbers of sensors and instrumentations such as an angular index table, fixtures, *etc*. Hence, there is a need for developing a suitable measurement and evaluation technique of spindle runout for understanding the machining performance of the machine tool.

Article history: received on Sep. 14, 2016; accepted on Dec. 04, 2016; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0018.

Laser-based optical measurement techniques have been developed by the researchers for evaluation of the spindle accuracy in machine tools. An optical measurement system consisting of a laser diode and position-sensitive detectors is used for measuring the spindle error during motions in high-speed conditions [4]. A laser interferometer is used for measuring the spindle rotation errors such as the radial motion error and axial motion error in a lathe [5]. An optical measurement system consisting of a rod lens, a ball lens, a laser beam, and a photodiode is developed for measuring rotational errors of a micro-spindle [6]. Fujimaki and Mitsui developed an optical measurement system consisting of a laser diode, a quadrant photodetector and a beam splitter for measuring the spindle radial runout of a miniaturized machine tool [7]. Though the laser-based measurement techniques have a longer working distance, they require extensive experimental arrangements and more setup time for aligning the laser path with the optics. With the recent advancements in computing and imaging systems, machine vision systems have been widely applied for different industrial inspection applications. A vision-based measurement technique using a CCD camera and a lens arrangement was proposed for measuring the radial errors of a cutting tool [8]. A change in position of the cutting tool is measured using a thresholding-based edge detection method.

It was noticed that the accuracy of spindle error measurement using a machine vision system was limited by the edge detection algorithms and lighting conditions. Hence, there is a need for developing suitable image processing algorithms to improve the accuracy of edge detection for estimation of the spindle runout using a machine vision system. Also, existing methods of spindle runout estimation are not suitable for on-machine inspection due to the requirements of multiple sensors and measurement setups for removing the contribution of the form error of the master cylinder. In order to overcome this difficulty, this work focuses on developing an image processing method suitable for online estimation of spindle runout in a lathe using a *Circular Hough Transform* (CHT)-based subpixel circle detection method. In the proposed method, a circle is detected in the images for measuring the radial error of the spindle; hence, it does not take into account the contribution of the form error of the master cylinder. The experimental results of the proposed method for evaluating the spindle radial error of a lathe are presented and discussed in this paper.

## 2. Development of machine vision system for spindle runout estimation

In the present work, a machine vision system consisting of a CMOS camera, a frame grabber and a PC image acquisition system is developed for estimation of the spindle radial error in a lathe. A precisely machined master cylinder is mounted on the lathe spindle and used as a target to measure the runout of the spindle. It is important to capture high quality images in the uniformly illuminated area of interest for machine-vision-based inspection applications. In the present work, a front lighting system with a ring arrangement of red LEDs is used to illuminate the circular face of the master cylinder. This lighting arrangement provides a shadow-free illumination and the red LED light is intensive enough to block the ambient light on the master cylinder. A lighting intensity is manually adjusted and controlled to provide uniform illumination on the circular face of the master cylinder using a regulated power supply. Details of the experimental arrangement for the spindle radial error estimation in a lathe are explained in this Section.

### *2.1. Experimental arrangement for image acquisition*

The machine vision system used for measurement of the spindle radial error consists of a monochrome CMOS camera (AVT Marlin F-131b), a frame grabber (IEEE-1394A) and a PC with the LABVIEW software (Version.8.0) for storing images of the spindle as shown

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

in Fig. 1a. General specifications of the CMOS camera used in the present work are listed in Table 1.

Table. 1 Specifications of the camera used for measurement of the spindle radial error.

| Items | Description |
|---|---|
| Camera Model | AVT Marlin F-131b |
| Image device | Type 2/3 (diag. 11 mm) global shutter CMOS sensor |
| Effective picture elements | 1280 (H) x 1024 (V) |
| Cell size | 6.7 µm x 6.7 µm |
| Resolution depth | 8 bit; 10 bit (ADC) |
| Lens mount | C-Mount |
| Digital interface | IEEE 1394 IIDC v. 1.3 |
| Power consumption | Less than 3 watt (@ 12 V DC) |
| Dimensions | 72 mm x 44 mm x 29 mm (L x W x H); w/o tripod and lens |

In order to measure the radial error of the spindle, a cylindrical master cylinder of 13 mm diameter is mounted on the lathe spindle and the CMOS camera is placed firmly on the tool post of the lathe to focus on the circular face of master cylinder. A distance between the camera and the master cylinder was measured using a standard scale and it was found to be 40 cm. The horizontal and vertical tilts of the camera in relation to the base of the tool post was checked using a spirit level, as shown in Fig. 1c. Screws in the tool post were manually adjusted until the bubble in the spirit level remained in the centre position, thus eliminating the misalignment of the camera.

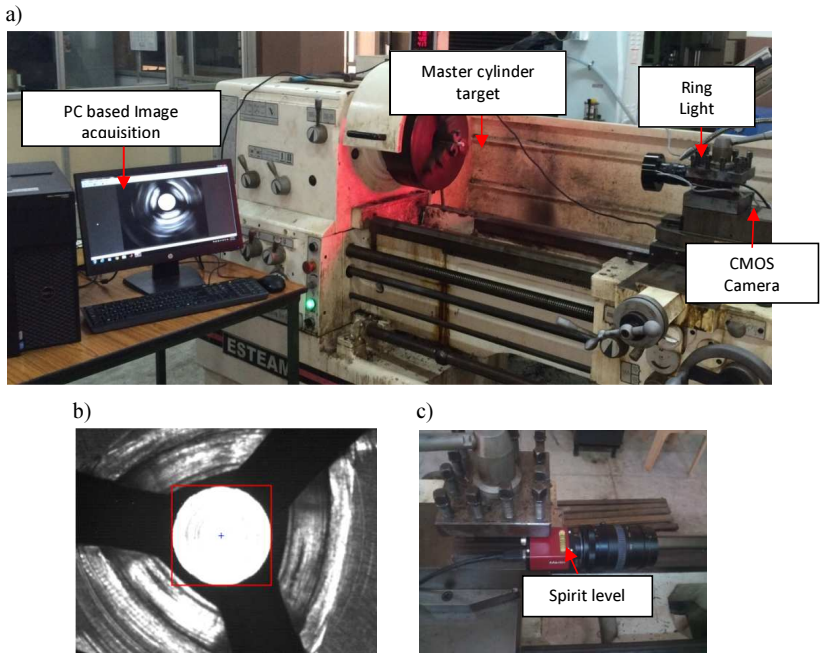



Fig. 1. The experimental arrangement for spindle radial error measurement using the vision system in a lathe. Important elements of the machine vision system for spindle radial error measurement (a); an image of the master cylinder (b); verification of alignment of the camera using a spirit level.

Further, the effect of misalignment of camera was analysed by calculating an aspect ratio of the circular face of the master cylinder in the image. The aspect ratio is defined as a ratio of the width of minimum enclosing rectangle of an object and the length of that object [9], as given below:

$$A = \frac{W}{L}. \tag{1}$$

Figure 1b shows the minimum enclosing rectangle for the master cylinder and the value of aspect ratio is calculated to be 1, which ensures the proper alignment of the camera. After verifying the alignment of the camera, a sequence of images of the master cylinder are captured with a resolution of 800 pixels x 600 pixels for different spindle speeds and stored using the LABVIEW Image acquisition software in the PC. As the maximum frame rate of the camera is 30 fps, the spindle radial error measurements were carried out at lower spindle speeds. Fig. 2 shows sample images acquired for a spindle speed of 25 rpm.

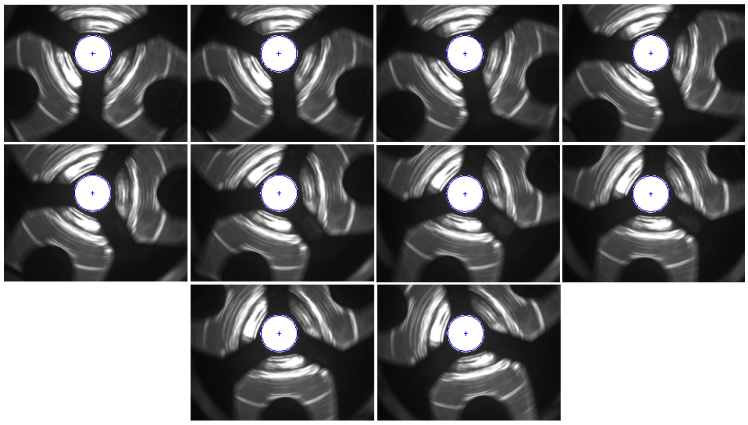


Fig. 2. A sequence of master cylinder images acquired at a spindle speed of 25 rpm.

Evaluation of radial error of the spindle using the digital images requires a suitable edge detection algorithm for detecting the change in position of the master cylinder, and calibration of the camera for specifying the measured values in the real world units.

### *2.2. Camera calibration*

The camera calibration is an essential step in machine vision inspection applications to obtain metric information from the images. In this work, the camera calibration is carried out using a standard slip gauge at a known distance [10]. Back lighting is used for acquiring the exact boundary of the slip gauge, as shown in Fig. 3. The number of pixels in *x* and *y* directions was counted in the image of the slip gauge and the scale factor for converting the pixel values into the real world units is determined using the dimension of the slip gauge in x, y directions, as given below:

Figure 3a shows the arrangement for acquiring the image of the slip gauge using a backlighting system. In this work, a slip gauge of dimension 30 mm x 4 mm is used, as shown in Fig. 3b; the numbers of pixels in *x*, *y* directions are found to be 361 pixels *x* 48 pixels, respectively. Hence, the conversion factor for obtaining measurements in the real world unit is calculated as 0.083 mm/pixel.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

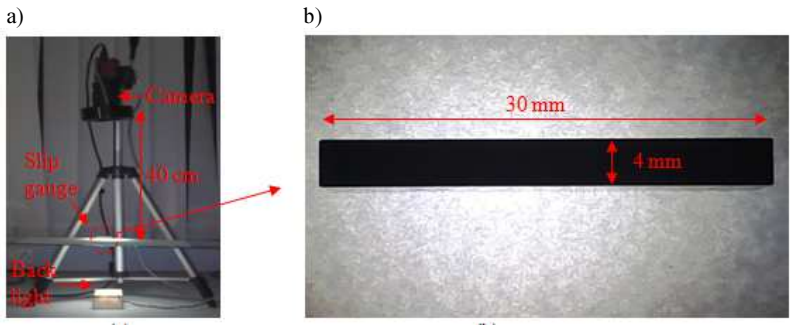



Fig. 3. Calibration of the camera using a slip gauge.

$$P_x = \frac{\text{Dimension of blocks in x direction}}{\text{Number of pixels in x direction}}, \tag{2}$$

$$P_y = \frac{\text{Dimension of blocks in y direction}}{\text{Number of pixels in y direction}}. \tag{3}$$

### *2.3. Edge detection using canny edge detection method*

The canny edge detection is a popular method for identifying the edge pixels of objects in the acquired images [11]. Commonly, edges in the digital images are detected based on significant changes in the grey level of pixels using first derivatives in respective directions. In the canny edge detection method, a magnitude of gradient and a direction of pixels are calculated for detecting changes in the grey level of pixels. The magnitude and direction of a gradient *G* are given by:

$$|\nabla \mathrm{I}| = G = \sqrt{Gx^2 + Gy^2}, \tag{4}$$

$$\theta = \operatorname{atan}\left( G_x, G_y \right), \tag{5}$$

where *Gx, Gy* are partial derivatives of the image *I* along x and y, respectively. The pixels with the gradient value above a threshold have been grouped and identified as edge pixels, and the remaining gradients below the threshold are lumped into the background with no information. Fig. 4 shows the results of edge detection using the canny edge detection method for the acquired image.

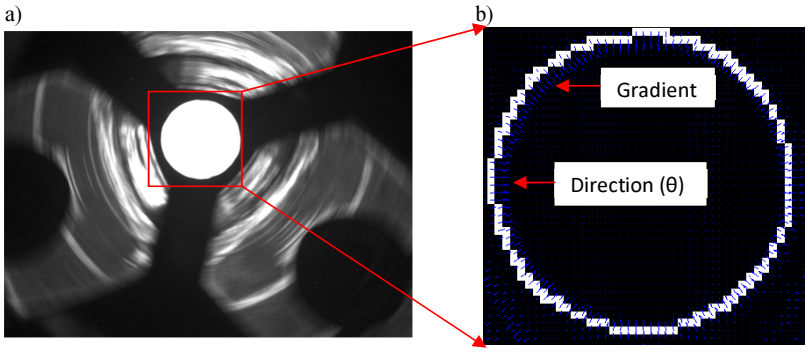



Fig. 4. Application of the canny edge detection for identifying edge pixels.
A grey scale image of the master cylinder (a); detected edge pixels using gradient (b).

In the presence of noise, the edge pixels identified by the canny edge detecting algorithm using gradients cannot define the boundary of the master cylinder accurately [10]. Hence, in the present work, the CHT is applied to the images for the accurate edge detection of the master cylinder and for evaluating the radial error of the spindle.

## 3. Circle detection using Circular Hough Transform

In the present work, the Circular Hough Transform is applied to the edge detection of master cylinder at a subpixel level to find the radial error of the spindle. The major advantage of this transform is its robustness towards irregularities in detected objects and disturbances like noise under varying illumination [12]. In the proposed method, the contribution of the form error of master cylinder is not taken into account, assuming the shape of the master cylinder to be a circle, for improving the accuracy of spindle radial error evaluation. The CHT is used to determine the circle parameters when the edge pixels are known. The steps involved in the CHT for the spindle runout estimation are shown in Fig. 5 and are explained in the subsequent Sections:

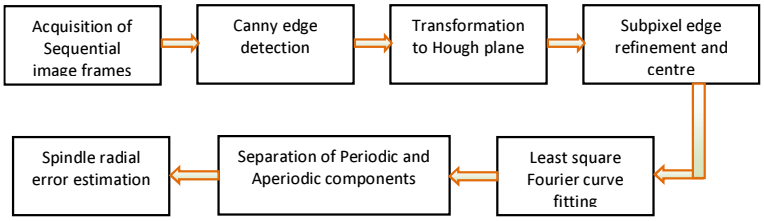



Fig. 5. The proposed method for spindle radial error evaluation.

### *3.1. Transformation of edge pixels for circle detection in Hough plane*

The key idea of the CHT is computation of the circle parameters [13], such as a circle centre and its radius ($X_c$, $Y_c$, $R$) in images by mapping the edge pixels in the image space onto the parameter space or the Hough space. The characteristic equation of a circle with a radius $R$ and centre ($X_c$, $Y_c$) is given below:

$$(x - X_c)^2 + (y - Y_c)^2 = R^2. \tag{6}$$

Here, the unknown parameters are the centre point's coordinates ($X_c$, $Y_c$) and the radius $R$. ($x$, $y$) is the edge location of a circle obtained by finding the maximum gradient above a predefined threshold value. Fig. 6 shows the transformation of an edge point in the image plane as the centre point of a circle with an unknown arbitrary radius $R$ in the Hough space.

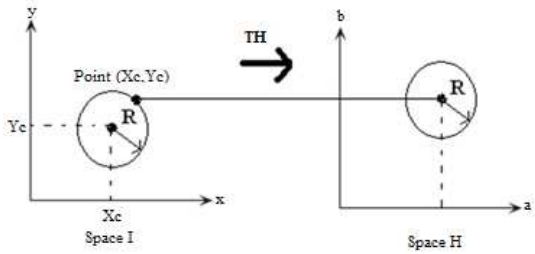



Fig. 6. Transformation of an edge point of a circle [14].

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

Each edge pixel of the image plane ($x$, $y$) is transformed onto the centre coordinates of a circle and the circle is drawn with the fixed radius ($R$) in the Hough plane using the gradient direction of the edge pixels, as given in [14]:

$$X_c = x - R * \cos(\theta), \tag{7}$$

$$Y_c = y - R * \sin(\theta), \tag{8}$$

where ($x$, $y$) are the locations of edge pixels obtained from the gradient and $\theta$ are the directions of the gradients of edge pixels. When this transformation is applied to all the edge pixels, it corresponds to the number of circles with a given arbitrary radius $R$ in the Hough plane, as shown in Fig. 7.

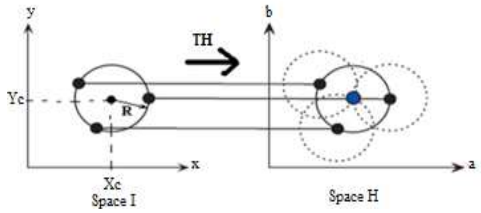



Fig. 7. Intersection of circles and centre estimation [14].

It is found that the edge pixels in the image plane form full circles with a desired radius $R$ in the Hough plane, where their intersection is identified as the centre point ($X_c$, $Y_c$) of the detected circle in the image plane.

### *3.2. Discretization of circle parameters and accumulator array computation*

The most important parameter while detecting a circle is its radius. It determines the size of circles plotted in the parameter space. In order to find the unknown radius of the circle in the image plane, a range of values ($R_{max}$, $R_{min}$) is chosen arbitrarily using the following constraint:

$$R = R_{min^2} < (x^2 + y^2) < R_{max^2}. \tag{9}$$

An accumulator array is initialized to count and store values of centre coordinates of the circles for all edge pixels and a given value of radius in the range of values ($R_{max}$, $R_{min}$), as below:

$$X_c = x - [R_{min}: R_{max}] * \cos(\theta), \tag{10}$$

$$Y_c = y - [R_{min}: R_{max}] * \sin(\theta). \tag{11}$$

For a given edge point in the circle of the image plane, if the circle is drawn with the desired radius in the Hough plane, the accumulator stores corresponding coordinates of the circle centre and radius. This count is increased for all the edge pixels in the accumulator array every time the circle is drawn with the desired radius of the circle. The accumulator array which provides the maximum count for coordinates of the circle centre and its radius is identified by a search method to find the circle centre and radius in the image plane at a subpixel level. The location of circle centre defines the location of datum axis of the master cylinder and detecting the edges of the master cylinder in the image plane.

### *3.3. Fourier series analysis of circle centre coordinates*

The estimated circle centres contain the contribution of the centring error of the spindle which is periodic in nature for every revolution of the spindle [3]. In the present work,

the periodic components of the circle centre coordinates are extracted using the Fourier curve fitting method in the time domain. The proposed mathematical model for interpreting the time sampled centre coordinates of the master cylinder is given by the following Fourier series formula:

$$X_{ci} = r_0 + \sum_{h=1}^{H}( a_h \cos(w * h * t_i) + b_h \sin(w * h * t_i)), \tag{12}$$

where $i$ = 1,2,3…$m$; and $m$ = the number of samples of centre coordinates considered for analysis; $H$ is the number of harmonics; $h$ = 1,2,…$H$; and ($a_h$, $b_h$) are Fourier coefficients which describe the repeatable components of the measurement data, such as the centring error, the form error of the target object, and the synchronous radial error of the spindle. $X_{ci}$ are the centre coordinates of master cylinder along $x$-axis, $X_{ci} = [X_{c1}, X_{c2}, \dots X_{cm}]^T$. The sampling time is calculated based on the frame rate of image acquisition and it is given by $t_i = [t_1, t_2, \dots t_m]^T$.

In the present work, the time taken for acquisition of an image frame is 1/30 of a second.

The time taken to complete one revolution is calculated from the time sampled circle centre data and it is denoted as $T$:

$$w = 2 * \frac{\pi}{T}. \tag{13}$$

A linear least square method is used to estimate the unknown Fourier coefficients by minimizing the sum of squares of deviations of measured data [15]:

$$\begin{bmatrix} X_{c1}' \\ X_{c2}' \\ X_{c3}' \\ . \\ . \\ X_{cm}' \end{bmatrix} = \begin{bmatrix} 1 & \cos(w*t_1) & \sin(w*t_1) & . & \cos(w*H*t_1) & \sin(w*H*t_1) \\ 1 & \cos(w*t_2) & \sin(w*t_2) & . & \cos(w*H*t_2) & \sin(w*H*t_2) \\ 1 & \cos(w*t_3) & \sin(w*t_3) & . & \cos(w*H*t_3) & \sin(w*H*t_3) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & \cos(w*t_m) & \sin(w*t_m) & . & \cos(w*H*t_m) & \sin(w*H*t_m) \end{bmatrix} \begin{bmatrix} r_0 \\ a_1 \\ b_1 \\ . \\ . \\ a_H \\ b_H \end{bmatrix}. \tag{14}$$

The above equation can be simplified to:

$$D = \begin{bmatrix} 1 & \cos(w*t_1) & \sin(w*t_1) & . & \cos(w*H*t_1) & \sin(w*H*t_1) \\ 1 & \cos(w*t_2) & \sin(w*t_2) & . & \cos(w*H*t_2) & \sin(w*H*t_2) \\ 1 & \cos(w*t_3) & \sin(w*t_3) & . & \cos(w*H*t_3) & \sin(w*H*t_3) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & \cos(w*t_m) & \sin(w*t_m) & . & \cos(w*H*t_m) & \sin(w*H*t_m) \end{bmatrix}$$

$$x = [r_0, a_1, b_1 \dots a_H, b_H]^T$$

Therefore:

$$X_c^{'} = Dx. \tag{15}$$

The Equation (14) leads to an over-determined system of simultaneous linear equations (*i.e.* $m > 2H + 1$). In this case, there exist residuals between the measurement data and the fitted curve, given by:

$$e_i = (X_{ci}' - Dx). \tag{16}$$

Assuming the residuals follow a normal probability distribution, the solution for the unknown model parameters can be obtained by minimizing the sum of squared residuals using a linear least square approach, as given by (17):

$$\hat{x} = [((D^T D)^{-1} D^T] X_{ci}' = [\hat{r}_0, \hat{a}_1, \hat{b}_1, \dots \hat{a}_H, \hat{b}_H]. \tag{17}$$

Here, $H$ represents the number of harmonics considered for evaluating the radial error of the spindle. The centring error of the master cylinder represents the first harmonic ($h$ = 1) and it can be removed from the measurement data, which is given as:

$$\hat{X}_{c\_cen} = \hat{a}_1 \cos(w * t_i) + \hat{b}_1 \sin(w * t_i). \tag{18}$$

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

The remaining harmonic components ($H > 2$) contribute to the synchronous radial error of the spindle and they can be extracted using the following formula:

$$\hat{X}_{c\_syn} = \sum_{h=2}^{H} \hat{a}_h \cos(w * H * t_i) + \hat{b}_h \sin(w * H * t_i). \tag{19}$$

This value is further analysed in a polar plot for evaluating the synchronous radial error of the spindle. The residuals of the measurement data for the fitted curve represent the asynchronous radial error of the spindle which is calculated using (16). It is further analysed in the polar plot for evaluating the asynchronous radial error of the spindle.

### *3.4. Estimation of radial error of spindle in polar plot*

A polar plot is commonly used for displaying the spindle error evaluation results with a base circle [3] and it requires the angular position of the spindle. The angular position of the spindle for a given time $t_i$ can be calculated using the following formula:

$$\theta_i = \omega * t_i. \tag{20}$$

Here, the value of $w$ is calculated using (13). The above equation is useful in plotting the synchronous and asynchronous radial error values of the spindle in a polar plot.

In accordance with the *ANSI/ASME B89.3M* standard, the least squares circle centre is calculated from the periodic components used for evaluating the synchronous radial error of the spindle after removing the contribution of the centring error of the spindle[16]. The asynchronous radial error is calculated from the aperiodic components of circle centre as the maximum deviation for a given spindle speed.

## 4. Results and discussions

The CHT-based circle detection approach is applied to a sequence images for estimation of the master cylinder centre. The results for estimated values of circle centres are presented for the sequence of images obtained for a spindle speed of 25 rpm. Further, the centre coordinates of the master cylinder are analysed using the least squares curve fitting technique to separate the contribution of the centring error of the master cylinder and the synchronous and asynchronous errors of the spindle. As the least squares curve fitting method provides an approximation of the ideal curve assuming the residuals follow a normal distribution, the error of the estimation obtained by the least squares curve fitting method is evaluated using the simulated circle centre data. The simulation and experimental results of the least squares curve fitting method are presented. Further, the spindle radial error values are evaluated for different spindle speeds and the results are presented.

### *4.1. Estimation circle centre using CHT*

In order to reduce the computation time and complexity of the transform, a range of radii of the master cylinder has been fixed manually. Fig. 8 shows the accumulator array computation results in the Hough space in 2D and 3D views for different ranges of radii. When a broader radius range of 15 pixels ($R_{max} - R_{min} = 15$) is fixed, the maximum votes for the centre of the circle accumulate at (386, 214), as shown in Fig. 8a.

To further reduce the computation time, a finer range of 6 pixels is fixed and the computed accumulator array is shown in Fig. 8b. In this case, the circle centre estimation is also found to be (386,214). This proves the robustness of the CHT method in detecting the circle centre coordinates in the image plane for a change of the search radius in the image plane. The estimated centre coordinates of circle in the Hough plane are located in the image plane

and they are used in identifying the edge of the master cylinder at a subpixel level, as shown in Fig. 9. The subpixel level identification of circle edge is shown in Fig. 9b as compared with the pixel level edge detection using a canny edge detector. This result confirms an improved edge detection at a subpixel level for the master cylinder in given images as compared with the conventional canny edge detection method. The estimated values of circle centre require further analysis in the time domain for evaluation of the radial error of the spindle.

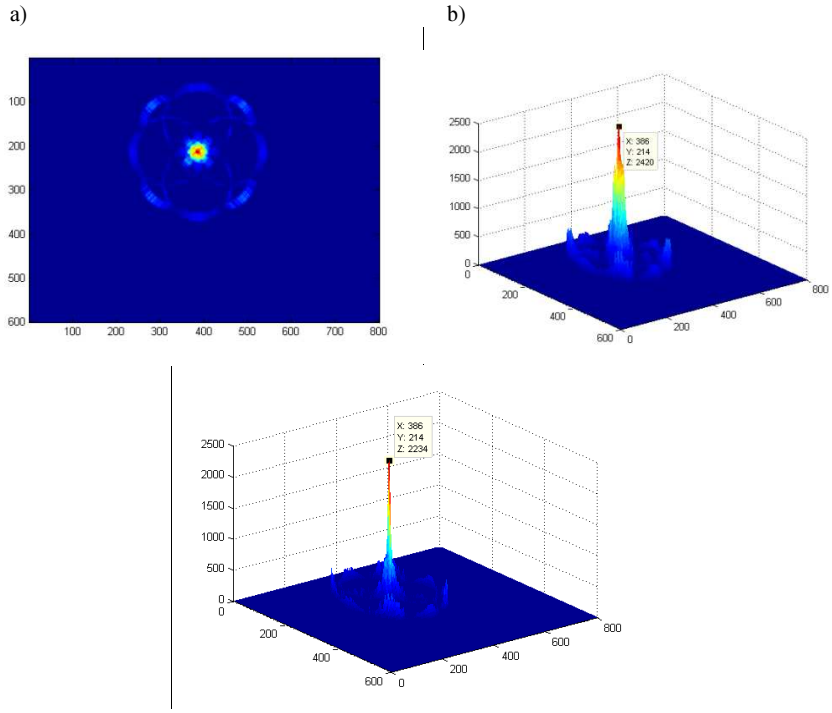



Fig. 8. 2D and 3D views of the accumulator array for the computation of circle centre. A larger width of radius range (15 pixels) (a); a smaller width of radius range (6 pixels) (b).

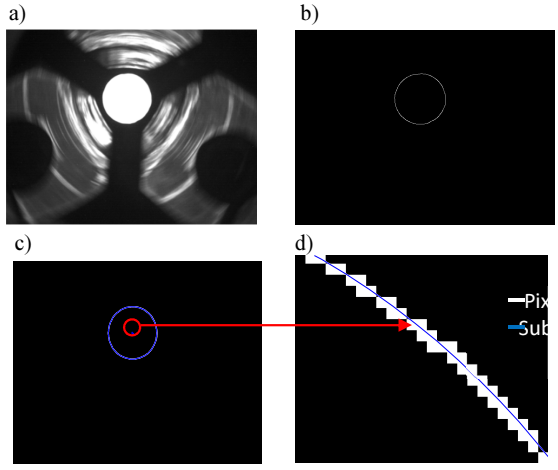



Fig. 9. Pixel and Subpixel level edge refinement. Input image (a); canny edge (b); Hough circle fitting (c); subpixel edge (d).

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

### *4.2. Time domain analysis of circle centre data*

The CHT is applied and estimated The values of circle centre estimated with applying CHT to a sequence of images for different spindle speeds are shown in Fig. 10. They indicate that a change in position of the master cylinder in the Cartesian plane is found to remain within a range of 212.5–214.5 pixels in Y direction and 385−387.5 pixels in X direction.

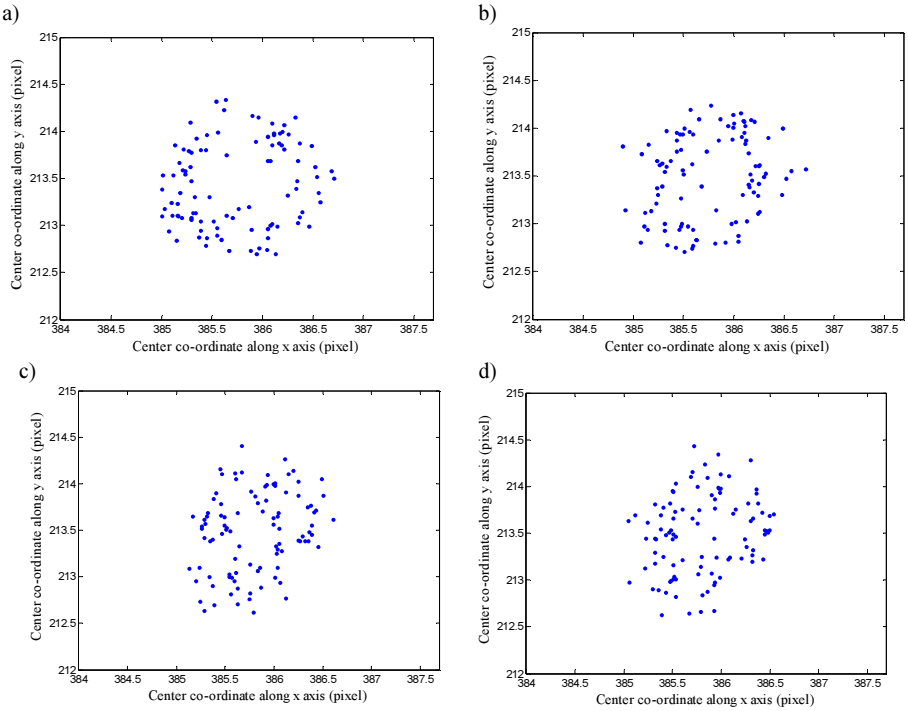



Fig. 10. The values of circle centre coordinates estimated using CHT.
25 rpm (a); 50 rpm (b); 75 rpm (c); 100 rpm (d).

Evaluation of radial error of the spindle requires further analysis of the circle centre values in the time domain. As the frame rate of camera is 30 frames/sec, a time stamp is attached to the circle centre coordinates for a given image frame. Table 2 shows samples of the circle centre coordinates and the sampling time for a spindle speed of 25 rpm.

Table 2. Samples of the circle centre coordinates estimated using the CHT method.

| Frame number | Time (Sec) | Coordinates of circle centre (pixels) | |
|---|---|---|---|
| | | *Xc* | *Yc* |
| Frame 1 | 0.00 | 386.2496 | 213.3237 |
| Frame 2 | 1/30 | 385.7641 | 213.182 |
| Frame 3 | 2/30 | 385.9357 | 212.6999 |
| Frame 4 | 3/30 | 386.0461 | 212.8709 |
| Frame 5 | 4/30 | 386.0446 | 212.7411 |
| Frame 6 | 5/30 | 385.5476 | 212.9786 |
| Frame 7 | 6/30 | 385.4458 | 212.8686 |
| Frame 8 | 7/30 | 385.5866 | 212.8536 |
| Frame 9 | 8/30 | 385.3831 | 212.9457 |
| Frame 10 | 9/30 | 385.2956 | 213.0832 |

The mean value of the circle centre along X direction is calculated and subtracted from each centre coordinate to provide a reference in the time domain. Further, the units of the centre coordinates in the images are converted from pixels to microns using a calibration value of 83 micron/pixel. Fig. 11 shows samples of corrected and calibrated mean values of the circle centre for a spindle speed of 25 rpm.

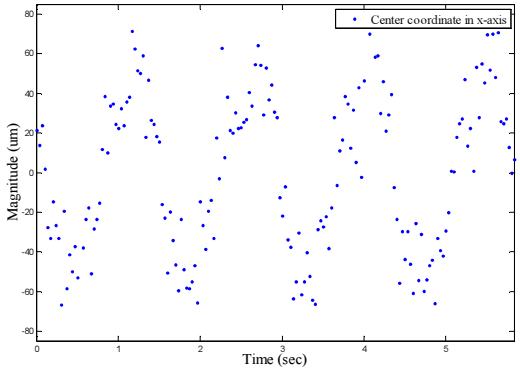



Fig. 11. The corrected and calibrated mean coordinates of circle centre of the master cylinder for a spindle speed of 25 rpm.

A periodically varying sinusoidal trend is observed in the circle centre data in the time domain and it is due to the combined contribution of centring of master cylinder and errors in the axis of rotation of the spindle. The centring error of the master cylinder is considered as a systematic error since it is due to inaccurate mounting of the master cylinder in the spindle [3]. Hence, to evaluate the radial error of the spindle, its contribution needs to be removed. In the present work, to remove the contribution of centring error of the master cylinder, the Fourier curve fitting method is applied to the circle centre data.

### *4.3. Simulation of circle centre data and application of least squares curve fitting method*

A Fourier harmonic series given by the equation (12) is used for generating the periodic components of circle centre data for typical values of model coefficients. Table 3 shows the assumed model coefficients applied to characterizing the periodic components of circle centre data using the first 5 harmonics. Here, the number of harmonics is limited to 5 and a magnitude of the first harmonic is assumed to be higher following thesinusoidal trend in the experimental data and it contributes the centring error of the master cylinder. Magnitudes of other harmonics ($H > 2$) are assumed based on typical values obtained in the experimental data. Further, the asynchronous components of circle centre data are assumed to follow a normal probability distribution with a given standard deviation of 0.5 pixels. The synchronous and asynchronous values of circle centre data are combined to provide the simulated circle centre data.

The least squares curve fitting method is applied to the simulated circle centre data to decompose the periodic and aperiodic components. Fig. 12 shows the curve fitted to the simulated data and following the general sinusoidal trend.

The values of harmonic components estimated using the least square curve fitting method are shown in Table 3. It can be seen that deviations between the estimated and simulated values are found to be less than 2.47%. This proves the effectiveness of the least squares curve fitting method for analysing the circle centre data.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

Table 3. Comparison between the simulated and estimated values of model coefficients using the curve fitting method.

| Model Coefficients | Simulated values | Estimated values | Error (%) |
|---|---|---|---|
| $a_1$ | 5.5729 | 5.635 | 1.102 |
| $b_1$ | 47.3218 | 47.32 | 0.0038 |
| $a_2$ | 4.1323 | 4.115 | 0.4204 |
| $b_2$ | 2.991 | 2.989 | 0.0669 |
| $a_3$ | 4.0132 | 3.965 | 1.2156 |
| $b_3$ | 2.8978 | 2.94 | 1.4354 |
| $a_4$ | 1.0267 | 0.9513 | 7.926 |
| $b_4$ | 3.5621 | 3.697 | 3.6489 |
| $a_5$ | 2.0014 | 2.003 | 0.0799 |
| $b_5$ | 1.0745 | 1.065 | 0.892 |
| $r_0$ | 0.0405 | 0.04121 | 1.7229 |

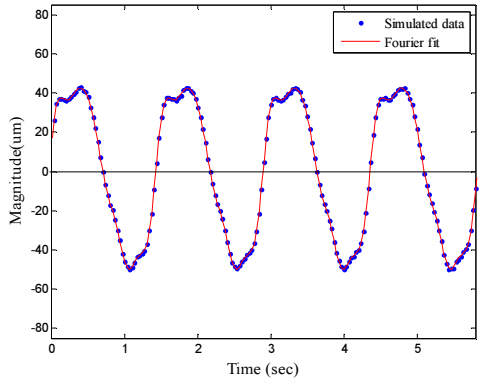



Fig. 12. Application of the least squares curve fitting method for the simulated circle centre data.

### *4.4. Experimental results of least squares curve fitting method*

In order to separate the contribution of the centring error and to evaluate the radial error of the spindle, the coordinates of the circle centre in X direction are further analysed in the time domain using the least squares curve fitting method. The estimated coordinates of circle centre are plotted in the time domain based on a frame rate of image acquisition.

Figure 12 shows the corrected and calibrated mean coordinates of circle centre of the master cylinder and they exhibit a sinusoidal trend which is due to the contribution of the centring error of the master cylinder. In order to remove this contribution and extract the radial error values of the spindle, a least square curve fitting algorithm is applied to circle centre data. Here, a harmonic cut-off value is selected as 15. It can be seen that the fitted curve closely follows the periodic trend of the circle centre data, as shown in Fig.13a. It is also noticed that the periodic trend is repetitive for each revolution of the spindle and the time taken for completion of one revolution is also computed using the periodic trend. The total indicated runout of the spindle is calculated as 142 microns. As shown in Fig. 13b, the residuals show random variations as the periodic components are extracted by the Fourier curve, and they represent the contribution of the asynchronous radial error of the spindle [17].

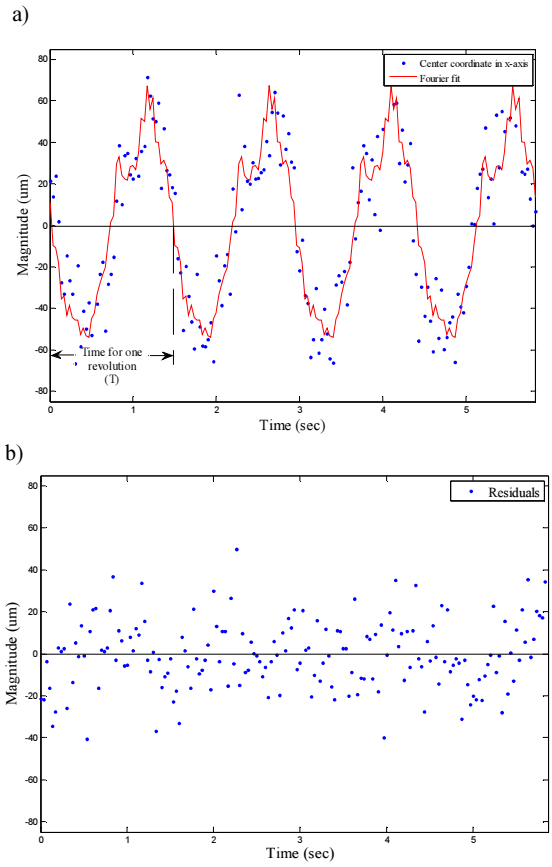



Fig. 13. Fourier series analysis of the circle centre data in X direction.
Extraction of periodic components using the least squares fitting of the Fourier curve (a); residuals (b).

The Fourier coefficients estimated by using the least squares curve fitting method is shown in Table 4 for the first five harmonic components. It can be noticed that the first hamonic components are dominant and they represent the contibution of the centering error of the master cylinder. Magnitudes of other harmonic components are found to be less than the first harmonic component one and they represent the combined contribution of the synchronous radial error of the spindle which is due to imperfections in the bearing surface of the spindle.

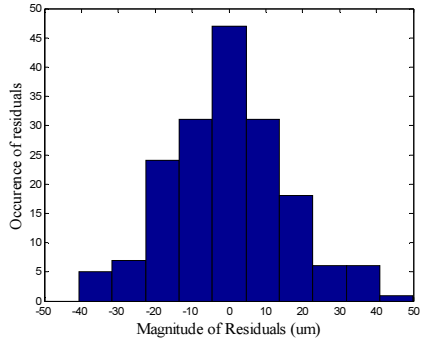



Fig. 14. A histogram of residuals.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

Table 4. The values of model parameters of the Fourier curve estimated using the least square method.

| Model Coefficients | Estimated values of model coefficients using the proposed method |
|---|---|
| $a_1$ | 32.3702 |
| $b_1$ | –39.4984 |
| $a_2$ | 8.4089 |
| $b_2$ | 1.5164 |
| $a_3$ | 6.7610 |
| $b_3$ | –3.7970 |
| $a_4$ | –2.1307 |
| $b_4$ | –2.5259 |
| $a_5$ | –2.8483 |
| $b_5$ | 2.2809 |
| $r_0$ | 0.0512 |

The residuals of the fitted curve from the circle centre data are analysed in a histogram to understand the nature of variation and they are showed in Fig. 14. It can be noticed that most of the residuals are centred around the zero mean which indicates that the distribution is represented by a bell curve. Also, these results validate the assumption about the residuals which follow a normal distribution for the least squares fitting of the Fourier curve.

### *4.5. Evaluation of spindle radial error*

In the present work, a basic circle radius of 5 units is used for displaying the separated components of circle centre data. It is shown in Fig. 15. These plots are obtained for the estimated angular position of the spindle using (20). It can be seen from Fig. 15a that the polar profile of the circle centre data deviates from the base circle, which is due to the contribution of the centring error of the master cylinder. The contribution of the centring error of the master cylinder is separated by using the first harmonic component of the fitted Fourier curve [18] using (18); it is shown in Fig. 15b. It indicates a clear deviation from the base circle and the polar chart centre, which is due to a misalignment of the master cylinder in the axis of rotation of the spindle.

The periodic components of the circle centre data after separation of the centring error is calculated using (19); it is given in Fig. 15c. As the synchronous components of radial error are periodic and repeatable, a magnitude of variation is significantly smaller for every revolution and its value obtained by using the least squares circle centre is 10.660 microns. This error provides a limiting value for the roundness error of the cylindrical components using the spindle.

The aperiodic components of circle centre data are analysed for evaluating the asynchronous radial error of the spindle; it is shown in Fig. 15d. The asynchronous error of the spindle includes the contribution from structural motion of the machine structure and it is found to be non-repeatable for every revolution. It is evaluated as the maximum deviation of aperiodic components and it is found to be 85.521 microns. This error value provides the baseline value for the surface finish of the components using the spindle.

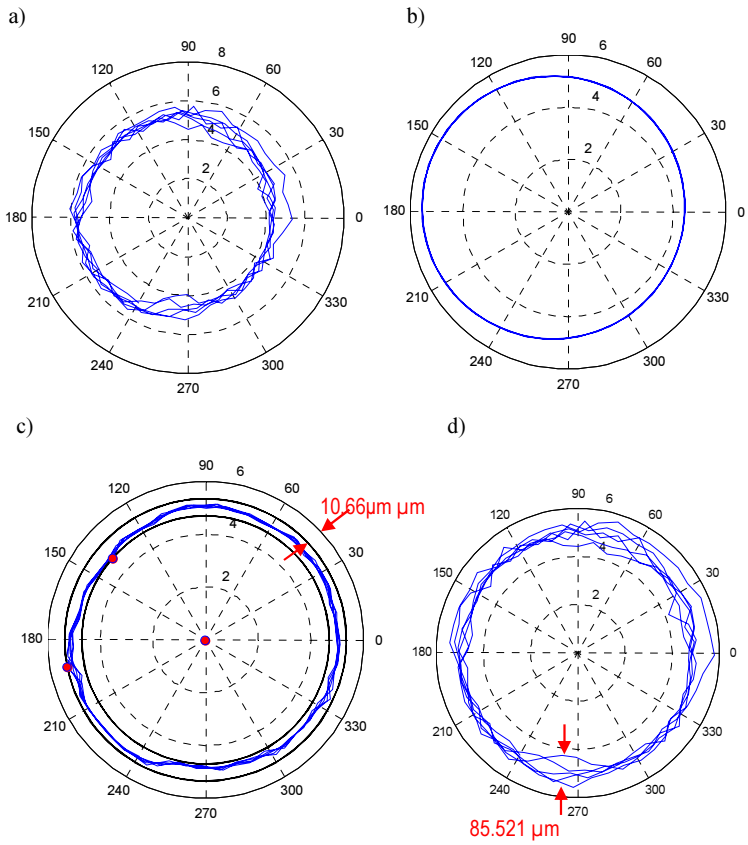



Fig. 15. Separation of the centring error and the spindle radial error for a spindle speed of 25 rpm. Coordinates of circle centre in X direction (a); centring error (b); the synchronous radial error (c); the asynchronous radial error (d).

The estimated values of synchronous and asynchronous radial errors provide a limiting value for the roundness error and surface finish of the components using the spindle. It can be noticed that the synchronous radial error is found to be decreasing with the increase in spindle speed due to a change in spindle contact during spindle rotation at the form variations in the bearing surfaces for a given spindle speed [15]. It can be noticed from Fig. 10 that the coordinates of circle centre move closer towards the average position of the master cylinder for the increase in spindle speed, hence there is a decrease in the synchronous radial error of the spindle. However, the asynchronous radial error value is found to vary randomly as it is aperiodic and also it includes the contribution from structural motion and vibration of the spindle which varies for different spindle speeds.

### *4.6. Repeatability of spindle radial evaluation using proposed method*

In order to understand the repeatability of the spindle radial error measurements using the proposed method, the experiments have been repeated for four times and standard deviations of the estimated values were calculated and are listed in Table 5 and Table 6. It can be noticed that there is no much deviation in magnitudes of the evaluated values of synchronous and asynchronous radial errors for different experiments and they are of a similar order at various spindle speeds.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

Table 5. Evaluation of the synchronous spindle radial error at different spindle speeds.

| Speed (rpm) | Evaluation of Synchronous Error (µm) | | | | Mean (µm) | Standard deviation (µm) |
|---|---|---|---|---|---|---|
| | Experiments | | | | | |
| | I | II | III | IV | | |
| **25** | 10.66 | 10.67 | 10.75 | 10.54 | 10.655 | ± 0.086603 |
| **50** | 13.164 | 13.152 | 13.162 | 13.158 | 13.159 | ± 0.005292 |
| **75** | 10.198 | 10.194 | 10.189 | 10.183 | 10.191 | ± 0.006481 |
| **100** | 7.542 | 7.538 | 7.541 | 7.537 | 7.5395 | ± 0.002380 |

Table 6. Evaluation of the asynchronous spindle radial error at different spindle speeds.

| Speed (rpm) | Evaluation of asynchronous Error (µm) | | | | Mean (µm) | Standard deviation (µm) |
|---|---|---|---|---|---|---|
| | Experiments | | | | | |
| | I | II | III | IV | | |
| **25** | 85.521 | 85.517 | 85.429 | 85.642 | 85.52725 | ± 0.087492 |
| **50** | 63.523 | 63.537 | 63.523 | 63.53 | 63.52825 | ± 0.006702 |
| **75** | 56.234 | 56.241 | 56.242 | 56.246 | 56.24075 | ± 0.004992 |
| **100** | 65.021 | 65.023 | 65.021 | 65.028 | 65.02325 | ± 0.003304 |

The maximum standard deviations of the evaluated synchronous and asynchronous radial errors are found to be ± 0.086603 µm and ± 0.087492 µm, respectively. These results prove the repeatability of the proposed method for evaluation of the spindle radial error at a submicron level.

### *4.7. Comparison with runout estimation using dial indicator*

A dial indicator is used to measure the runout of the master cylinder in a lathe, as shown in Fig. 16. The peak-to-peak variation in the dial was calculated as 150 microns for one revolution of the spindle. The proposed machine vision system provides estimation of total indicated runout of the spindle as 142 microns, as shown in Fig. 13a.

Fig. 16. Runout measurement using a dial indicator.

Since the runout estimation using a dial indicator includes the contribution from the centring error of the master cylinder, the form error of the master cylinder and the spindle radial error, its magnitude is found to be higher than the value estimated by the proposed machine vision system. That is because in the proposed method the contribution of the centring error and

the form error of the master cylinder is removed. Also, the synchronous and asynchronous radial errors are separately calculated using the curve fitting method at different spindle speeds.

## 5. Conclusions

Spindle radial error evaluation is an important task in understanding the machining capability of a machine tool's spindle. This paper demonstrates application of a machine vision system and a CHT-based image processing technique to evaluating the radial error of a lathe spindle at different spindle speeds. Application of CHT to detecting a circle in the image is found to be robust in estimating the circle parameters. Also, the circle detection method provides a simpler approach to eliminating the contribution of the form error of the master cylinder. In order to extract the contribution of the centring error and the radial error of the spindle, the periodic and aperiodic components of circle centre are analysed using the Fourier series curve fitting method. The synchronous radial error of a lathe spindle is found to vary between 7.542 microns and 13.164 microns for different spindle speeds and it showed a decreasing trend with the increase of the spindle speed. However, the asynchronous radial error value is found to vary randomly within a range of 56.234 microns – 85.521 microns, as it includes the contribution from structural motion of the spindle which varies for different spindle speeds. The repeatability of the spindle radial error evaluation using the proposed method is found to be at a submicron level. The proposed method can be extended to the online monitoring and estimating the spindle radial errors using a high-speed camera.

## Acknowledgements

The authors wish to thank Department of Science Technology, New Delhi for providing the necessary funding to establish the machine vision system for spindle radial error measurement in machine tools under the fast track Young scientist scheme.

## References

[1] Bryan, J.B., Vanherck, P. (1975). Unification of terminology concerning the error motion of axes of rotation. *California Univ., Livermore (USA). Lawrence Livermore Lab.*

[2] Murugarajan, A., Samuel, G.L. (2011). Measurement, modeling and evaluation of surface parameter using capacitive-sensor-based measurement system. *Metrol. Meas. Syst.*, 18(3), 403–418.

[3] Marsh, E.R. (2008). *Precision spindle metrology*. DEStech Publications.

[4] Liu, C.H., Jywe, W.Y., Lee, H.W. (2004). Development of a simple test device for spindle error measurement using a position sensitive detector. *Measurement science and Technology*, 15(9), 1733.

[5] Castro, H.F. (2008). A method for evaluating spindle rotation errors of machine tools using a laser interferometer. *Measurement*, 41(5), 526–37.

[6] Murakami, H., Kawagoishi, N., Kondo, E., Kodama, A. (2010). Optical technique to measure five-degree-of-freedom error motions for a high-speed microspindle. *International Journal of Precision Engineering and Manufacturing*, 11(6), 845–50.

[7] Fujimaki, K., Mitsui, K. (2007). Radial error measuring device based on auto-collimation for miniature ultra-high-speed spindles. *International Journal of Machine Tools and Manufacture*, 47(11), 1677–85.

[8] Deakyne, T.R., Marsh, E.R., Lehman, J., Bartlett, B., Solutions, C. (2008). Machine vision with spindle metrology using a ccd camera. *Proc. of ASPE 23nd Annual Meeting, American Society for Precision Engineering*, Portland.

[9] Chen, A., Liu, N. (2010). Circular object detection with mathematical morphology and geometric properties. *IEEE, International Conference on Computer, Mechatronics, Control and Electronic Engineering*, 2, 318–321.

*Metrol. Meas. Syst.*, Vol. 24 (2017), No. 1, pp. 201–219.

[10] Jain, A.K. (1989). *Fundamentals of digital image processing.* Prentice-Hall.

[11] Ding, L., Goshtasby, A. (2001). On the Canny edge detector. *Pattern Recognition*. 34(3), 721–725.

[12] Smereka, M., Dulęba, I. (2008). Circular object detection using a modified Hough transform. *International Journal of Applied Mathematics and Computer Science*, 18(1), 85–91.

[13] Shetty, P. (2011). Circle Detection in Images. *Faculty of San Diego State University*, 9–11.

[14] Rhody, H. (2005). Lecture 10: Hough circle transform. *Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology.*

[15] Ashok, S.D., Samuel, G.L. (2012). Modeling, measurement, and evaluation of spindle radial errors in a miniaturized machine tool. *The International Journal of Advanced Manufacturing Technology*, 59(5–8), 445–61.

[16] American National Standards Institute. (2010). ANSI / ASME Axes of Rotation: Methods for Specifying and Testing. American Society of Mechanical Engineers.

[17] Ashok, S.D., Samuel, G.L. (011). Least square curve fitting technique for processing time sampled high speed spindle data. *International Journal of Manufacturing Research,* 6(3), 256–276.

[18] Ashok, S.D., Samuel, G.L. (2012). Harmonic-analysis-based method for separation of form error during evaluation of high speed spindle radial errors. *Proc. of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 0954405411434868.

# CONTENTS OF THE VOLUME 23/2016 M&MS

## Number 1/2016



## Number 2/2016

**Number 3/2016**



**Number 4/2016**

# Instructions for Authors

## Types of contributions

The following types of papers are published in *Metrology and Measurement Systems*:

- invited review papers presenting the current stage of the knowledge (max. 20 edited pages, 3000 characters each),
- research papers reporting original scientific or technological advancements (10–12 pages),
- papers based on extended and updated contributions presented at scientific conferences (max. 12 pages),
- short notes, *i.e.* book reviews, conference reports, short news (max. 2 pages).

## Manuscript preparation

***The text*** of a manuscript should be written in clear and concise English. The form similar to "camera-ready" with an attached separate file – containing illustrations, tables and photographs – is preferred. For the details of the preferred format of the manuscripts, Authors should consult a recent issue of the journal or the ***sample article*** and the ***guidelines for manuscript preparation***. The text of a manuscript should be printed on A4 pages (with margins of 2.5 cm) using a font whose size is 12 pt for main text and 10 pt for the abstract; an **even number of pages** is strongly recommended. The main text of a paper can be divided into sections (numbered 1, 2, ...), subsections (numbered 1.1., 1.2., ...) and – if needed – paragraphs (numbered 1.1.1., 1.1.2., ...). The title page should include: manuscript title, Authors' names and affiliations with e-mail addresses. The corresponding Author should be identified by the symbol of an envelope and phone number. A concise abstract of approximately 100 words and with 3–5 keywords should accompany the main text.

***Illustrations***, photographs and tables provided in the camera-ready form, suitable for reproduction (which may include reduction) should be additionally submitted one per page, larger than final size. All illustrations should be clearly marked on the back with figure number and author's name. All figures are to have captions. The list of figures captions and table titles should be supplied on separate page. Illustrations must be produced in black ink on white paper or by computer technique using the laser printer with the resolution not lower than 300 dpi, preferably 600 dpi. The thickness of lines should be in the range 0.2–0.5 mm, in particular cases the range 0.1–1.0 mm will be accepted. Original photographs must be supplied as they are to be reproduced (*e.g.* black and white or colour). Photocopies of photographs are not acceptable.

***References*** should be inserted in the text in square brackets, *e.g.* [4]; their list numbered in citation order should appear at the end of the manuscript. The format of the references should be as follows: for a journal paper – surname(s) and initial(s) of author(s), year in brackets, title of the paper, journal name (in italics), volume, issue and page numbers. The exemplary format of the references is available at the sample article.

## Manuscript submission and processing

***Submission procedure.*** Manuscript should be submitted via Internet Editorial System (IES) – an online submission and peer review system *http://www.editorialsystem.com/mms*

In order to submit the manuscript via IES, the authors (first-time users) must create an author account to obtain a user ID and password required to enter the system. From the account you create, you will be able to monitor your submission and make subsequent submissions.

The submission of the manuscript in two files is preferred: "Paper File" containing the complete manuscript (with all figures and tables embedded in the text) and "Figures File" containing illustrations, photographs and tables. Both files should be sent in DOC and PDF format as well as. In the submission letter or on separate page in "Figures File", the full postal address, e-mail and phone numbers must be given for all co-authors. The corresponding Author should be identified.

***Copyright Transfer.*** The submission of a manuscript means that it has not been published previously in the same form, that it is not under consideration for publication elsewhere, and that – if accepted – it will not be published elsewhere. The Author hereby grants the Polish Academy of Sciences (the Journal Owner) the license for commercial use of the article according to the Open Access License which has to be signed before publication.

***Review and amendment procedures***. Each submitted manuscript is subject to a peer-review procedure, and the publication decision is based on reviewers' comments; if necessary, Authors may be invited to revise their manuscripts. On acceptance, manuscripts are subject to editorial amendment to suit the journal style.

An essential criterion for the evaluation of submitted manuscripts is their potential impact on the scientific community, measured by the number of repeated quotations. Such papers are preferred at the evaluation and publication stages.

***Proofs***. Proofs will be sent to the corresponding Author by e-mail and should be returned within 48 hours of receipt.

**Other information**

***Author Benefits***. The publication in the journal is free of charge. A sample copy of the journal will be sent to the corresponding Author free of charge.

***Colour.*** For colour pages the Authors will be charged at the rate of 160 PLN or 80 EUR per page. The payment to the bank account of main distributor (given in "Subscription Information") must be acquitted before the date pointed to Authors by Editorial Office.

***Contact***:
***E-mail:*** *metrology@pg.gda.pl*
***URL***: *www.metrology.pg.gda.pl*
***Phone:*** (+48) 58 347-1357

***Post address***:
Editorial Office of *Metrology and Measurement Systems*
Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics
ul. Narutowicza 11/12, 80-233 Gdańsk, Poland