INDEX 330930 ISSN 0860-8229

M&MS

2017 4

QUARTERLY, VOLUME 24

METROLOGY AND MEASUREMENT SYSTEMS

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: 1.598 (5-Year Impact Factor 1.203).
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. 13,5 Ark. druk. 10,8
Papier offsetowy kl. III 80g 70 x 100 cm
Print run 120 copies

Druk: Centrum Poligrafii Sp. z o.o.
ul. Łopuszańska 53
02-232 Warszawa

www.czasopisma.pan.pl PAN www.journals.pan.pl

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 589–594.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

NOISE PROPERTIES OF GRAPHENE-POLYMER THICK-FILM RESISTORS

Krzysztof Mleczko[1)], Piotr Ptak[1)], Zbigniew Zawiślak[1)], Marcin Słoma[2)], Małgorzata Jakubowska[2)], Andrzej Kolek[1)]

1) Rzeszów University of Technology, Faculty of Electrical and Computer Engineering, W. Pola 2, 35-959 Rzeszów, Poland
(✉ kmleczko@prz.edu.pl, +48 17 865 1113, pptak@prz.edu.pl, zawislak@prz.edu.pl, akoleknd@prz.edu.pl)
2) Warsaw University of Technology, Faculty of Mechatronics, Św. A. Boboli 8, 02-525 Warsaw, Poland
(marcin.sloma@mchtr.pw.edu.pl, m.jakubowska@mchtr.pw.edu.pl)

Abstract

Graphene is a very promising material for potential applications in many fields. Since manufacturing technologies of graphene are still at the developing stage, low-frequency noise measurements as a tool for evaluating their quality is proposed. In this work, noise properties of polymer thick-film resistors with graphene nano-platelets as a functional phase are reported. The measurements were carried out in room temperature. $1/f$ noise caused by resistance fluctuations has been found to be the main component in the specimens. The parameter values describing noise intensity of the polymer thick-film specimens have been calculated and compared with the values obtained for other thick-film resistors and layers used in microelectronics. The studied polymer thick-film specimens exhibit rather poor noise properties, especially for the layers with a low content of the functional phase.

Keywords: graphene, polymer thick-film resistor, low-frequency noise, noise measurements.

1. Introduction

Graphene is a very promising material for potential applications in many fields, especially in micro- and nano-electronics. Its outstanding properties make it an ideal material for various applications, *e.g.* optoelectronics [1], RF communications [2], strain and pressure sensors [3, 4]. Due to its very high thermal conductivity, graphene is also used in electronic devices as a heat dissipation material [5]. One more application of graphene is its use as a conducting material in polymer resistors. *Polymer thick-film resistors* (PTFRs) have many advantages, among others a wide resistivity range, low processing temperature, low cost. Although they have been used in electronics for many years, they are considered as good candidates for the next generation of functional electronic components, especially due to their flexibility [6, 7]. Manufacturing technologies of this material are still in the developing stage. Therefore, many works dealing with PTFRs are focused on their electrical properties [8–10]. It is also well known that low-frequency noise measurements can be used as a tool for evaluating material quality. In this work, noise properties of polymer thick-film resistors, made of polymer and graphene, are studied in order to evaluate quality of the studied material.

2. Experiment

2.1. Specimens

Specimens for the measurements were manufactured in the thick-film technology. A resistive layer in PTFR specimen contains graphene nanoparticles dispersed in a polymer vehicle. Graphene nanoparticles were prepared from graphite using a modified Hummer's

Article history: received on Dec. 29, 2016; accepted on May 21, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0051.

method, acquired from Cheap Tubes Inc. Characteristic dimensions, estimated from *Scanning Electron Microscope* (SEM) observations, were 10 nm for average thickness and 15 µm for average particle diameter. The polymer vehicle selected to prepare the ink for printing was a solution of Mw 350 000 *poly-methyl metacrylate* (PMMA) in diethylene glycol butyl ether acetate (8 wt.%). Compositions of graphene nanoparticles in PMMA polymer vehicle were prepared with a modified mixing process used in thick-film material preparation. The main purpose of the mixing process is to prepare a well-dispersed paste without agglomerates. This was achieved by the sonication of carbon nanomaterials with dispersing agents in toluene for 60 minutes at room temperature. Malialim AKM-0531 dispersing agent provided by NOF Corporation was used for the surface treatment of carbon powders, to improve dispersion. Addition of 5 wt.% of the dispersing agent in respect to the weight of carbon fillers was sufficient to break agglomerates. After the partial evaporation of toluene, all specimens were mixed with PMMA vehicle in a mortar for 15 minutes. Afterwards, the pastes were homogenised on a three-roll mill with *silicon carbide* (SiC) rollers and a 5 µm gap. The specimens patterned into multi-terminal devices (see Fig. 2) of size 1 × 5 mm were printed with an AMI Presco 242 screen printer with 200 mesh stainless steel screens. Afterwards, the layers had been cured in 120ºC for one hour. Film thickness was estimated to be 10 µm. A SEM picture of cross-section of resistive layer after firing is shown in Fig. 1.

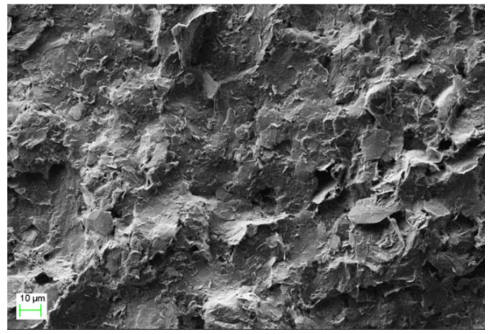

Fig. 1. A SEM picture of resistive layer.

2.2. Measurement setup

Low-frequency noise measurements have been performed in the dc Wheatstone bridge configuration. To dump fluctuations of the bias source voltage, a low-pass filter with a large time constant was placed between the dc source and the bridge. In the upper arms of the bridge, wire-wound resistors with a resistance much higher than that of specimens have been placed. The specimens were placed in the bottom arms of the bridge (Fig. 2). As thick-film resistors have been printed in pairs (on one substrate), two PTFR specimens of nearly the same resistance have been used. A noiseless wire-wound variable resistor has been connected in series with one of the PTFRs to enable balancing of the bridge. The amplified voltage signal from the bridge diagonal has been low-pass filtered and converted to digital samples in the *data acquisition* (DAQ) board. Then *power spectral densities* (PSDs) of voltage fluctuations S_{Vs} have been calculated.

The specimens have been mounted on a heat plate which temperature can be precisely controlled by an external temperature controller. The measurements were performed as a function of bias at room temperature.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 589–594.

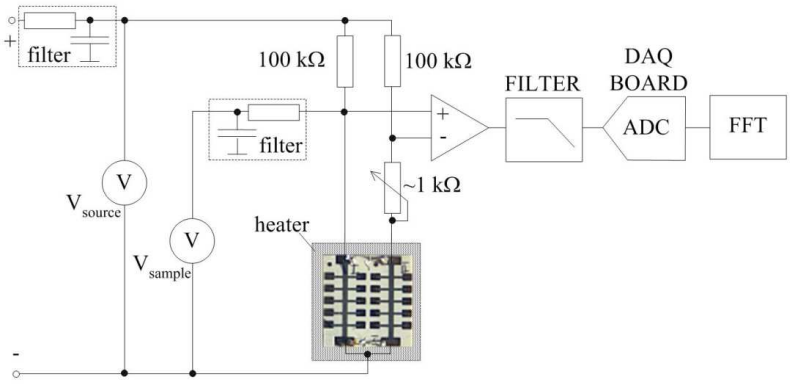

Fig. 2. The measurement setup.

3. Results

3.1. Noise intensity

In order to identify noise components, PSDs (S_{Vs}) of voltage fluctuations have been measured for several sample voltages, V_s. Then excess noise spectra have been calculated as $S_{Vex} = S_{Vs} - S_{Vs=0}$, where $S_{Vs=0}$ is the background noise, *i.e.* a PSD measured at $V_s = 0$. The resulted excess noise spectra measured at room temperature are plotted in Fig. 3.

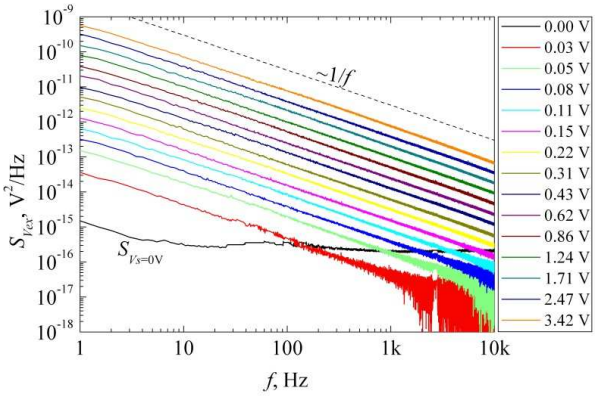

Fig. 3. The excess noise spectra, S_{Vex}, measured at room temperature for different bias voltages together with the background noise spectra, $S_{Vs=0}$.

Since the observed spectra $S_{Vex}(f)$ exhibit $1/f$ dependence, the product fS_{Vex} is frequency-independent. After averaging in some frequency band, Δf, it can be used as a reasonable measure of noise intensity. The data in Fig. 4 reveal that the noise intensity, $\langle fS_{Vex}\rangle_{\Delta f}$, scales linearly with the sample voltage squared. Usually, such a behaviour is interpreted as the evidence that $1/f$ noise is caused by the resistance fluctuations [11].

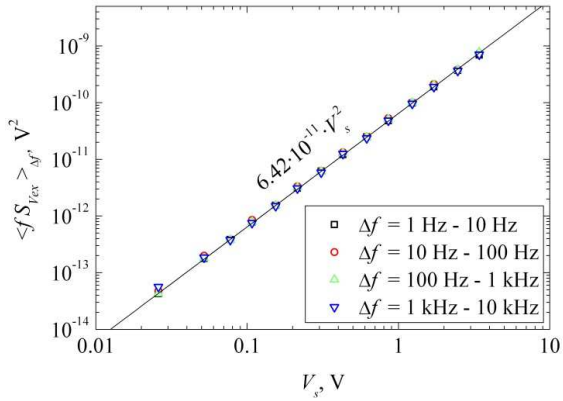

Fig. 4. The noise intensity, $\langle fS_{Vex} \rangle_{\Delta f}$, vs. sample voltage, V_s, for a PTFR specimen calculated for different frequency bands (points). The approximating line has a slope of 2.
The measurements were carried out at room temperature.

It stems from the observed scaling that the quantity $\langle fS_{Vex} \rangle_{\Delta f}/(V_s)^2$, which can be termed *relative noise intensity*, is voltage-independent and can be used to compare noise properties of different specimens, however of the same volume. This conditioning arises from the Hooge's phenomenological formula [12]:

$$S_V \sim \frac{V^2}{Nf}, \tag{1}$$

where N is the total number of carriers in a specimen which is proportional to the specimen volume.

3.2. Noise of resistive layer

In order to compare noise properties of different materials, a parameter $C \equiv$ volume$\cdot\langle fS_{Vex} \rangle_{\Delta f} V_s^{-2}$, independent of the specimen volume, should be used. The values of this parameter, evaluated for the studied PTFR specimen, are gathered in Table 1 and plotted in Fig. 5 as a function of specimen's resistivity. Another parameter, which is considered as a figure of merit in respect to $1/f$ noise and quality of the technology, is a ratio $K \equiv C/\rho$, where ρ is resistivity [13]. The parameter K is also independent of specimen's size and dimensions. The values of this parameter are also provided in Table 1 as well as in Fig. 5.

Table 1. The parameter values of the studied PTFR specimens.

Graphene content, wt %	ρ, Ω cm	K, μm^2/Ω	C, m^3
24	$2.43 \cdot 10^{-3}$	$3.13 \cdot 10^{-8}$	$7.61 \cdot 10^{-25}$
5	0.488	$2.87 \cdot 10^{-6}$	$1.40 \cdot 10^{-20}$
4	0.736	$1.39 \cdot 10^{-6}$	$1.02 \cdot 10^{-20}$

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 589–594.

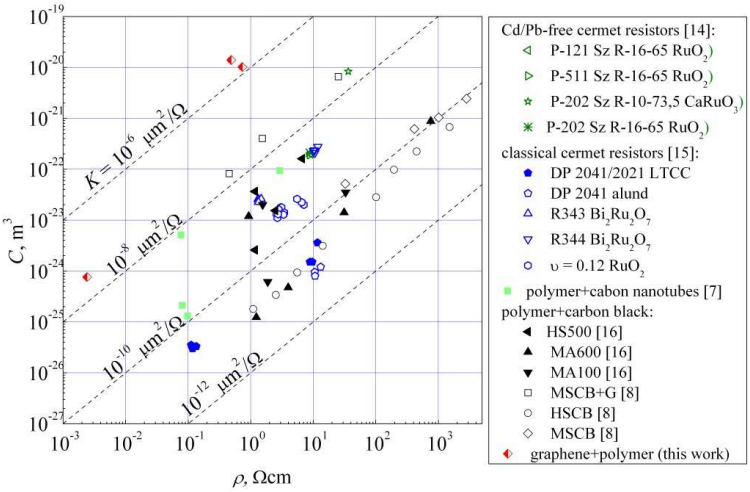

Fig. 5. The values of C and K parameters for different graphene-polymer thick-film resistors compared with those of other types of TFRs.

The values of C and K obtained for the studied PTFRs are compared with the values for other thick-film resistors used in microelectronics, especially those that use a polymer matrix. These values are presented in Table 2 and in Fig. 5. All of them are considerably lower. So one can conclude that noise properties of the studied specimens are rather poor and therefore the technology should be further improved, especially for the layers with a low content of the functional phase.

Table 2. The values of parameter K for various thick-film resistors (CNT – carbon nanotubes, MSCB – medium structure carbon black, HSCB – high structure carbon black).

Thick-film resistor type	K, μm²/Ω
RuO_2 on LTCC substrate [17]	$2.5 \cdot 10^{-11}$
RuO_2 on alumina substrate [17]	$4 \cdot 10^{-10}$
$Bi_2Ru_2O_7$ [17]	$2 \cdot 10^{-9}$
$CaRuO_3$ [15]	$2 \cdot 10^{-8}$
poly/CNT [7]	$3.3 \cdot 10^{-9}$
polymer+carbon black: MSCB, graphite+MSCB, HSCB [8]	$4.3 \cdot 10^{-11}$, $4 \cdot 10^{-8}$, $2.7 \cdot 10^{-11}$

4. Summary

1/f noise caused by resistance fluctuations has been found to be the main noise component in the studied specimens of polymer thick-film resistors. The values of noise parameter, either K or C, for PTFRs are substantially higher than those for other thick-film resistors used in microelectronics, especially for a low content of the functional phase. These observations might be helpful in further optimization of the technology process in order to gain a technological advantage.

Acknowledgements

This work was supported by the Rzeszow University of Technology, Department of Electronic Fundamentals Grant for Statutory Activity and statutory founding of IMiIB Warsaw University of Technology. The authors acknowledge the great contribution to the paper from late Prof. A. W. Stadler.

References

[1] Sensale-Rodriguez, B. (2015). Graphene-Based Optoelectronics. *J. Lightw. Technol.*, 33(5), 1100–1108.

[2] Palacios, T., Hsu, A., Wang, H. (2010). Applications of graphene devices in RF communications. *IEEE Communications Magazine*, 48(6), 122–128.

[3] Li, C., Gao, X., Guo, T., Xiao, J., Fan, S., Jin, W. (2015). Analyzing the applicability of miniature ultra-high sensitivity Fabry-Perot acoustic sensor using a nanothick graphene diaphragm. *Meas. Sci. Technol.*, 26(8), 085101.

[4] Smith, A.D., Vaziri, *et al.* (2013). Pressure sensors based on suspended graphene membranes. *Solid-State Electron.*, *88*, 89–94.

[5] Zhang, Y., Han, H., *et al.* (2015). Improved heat spreading performance of functionalized graphene in microelectronic device application. *Adv. Funct. Mater.*, 25(28), 4430–4435.

[6] Lostetter, A.B., Barlow, F., Elshabini, A., Olejniczak, K., Ang, S. (2000). Polymer thick film (PTF) and flex technologies for low cost power electronics packaging. *International Workshop on Integrated Power Packaging, IWIPP 2000*, 33–40.

[7] Słoma, M., Jakubowska, M., *et al.* (2011). Investigations on printed elastic resistors containing carbon nanotubes. *J. Mater. Sci. Mater. Electron.*, 22(9), 1321–1329.

[8] Dziedzic, A., Kolek, A. (1998). 1/*f* noise in polymer thick-film resistors. *J. Phys. D: Appl. Phys.*, 31, 2091–2097.

[9] Dziedzic, A. (2007). Carbon/polyesterimide thick-film resistive composites–experimental characterization and theoretical analysis of physicochemical, electrical and stability properties. *Microelectron. Rel.*, 47(2), 354–362.

[10] Srinivasa Rao, Y. (2007). Studies on electrical properties of polymer thick film resistors. *Microelectron. Int.*, 24(1), 8–14.

[11] Voss, R.F., Clarke, J. (1976). 1/*f* Noise from Systems in Thermal Equilibrium. *Phys. Rev. Lett.*, 36(1), 42–45.

[12] Hooge, F.N. (1976). 1/*f* noise. *Physica B*, 83, 14–23.

[13] Vandamme, L.K.J., Casier, H.J. (2004). The 1/f noise versus sheet resistance in poly-Si is similar to poly-SiGe resistors and Au-layers. *Proc. 34th European Solid-State Device Research Conf. 2004,* Leuven, Belgium. 365–368.

[14] Stadler, A.W., Kolek, A., *et al.* (2010). Noise properties of Pb/Cd-free thick film resistors. *J. Phys. D: Appl. Phys.*, 43(26), 265401.

[15] Stadler, A.W. (2011). Noise properties of thick-film resistors in extended temperature range. *Microelectron. Rel.* 51, 1264–1270.

[16] Fu, S.L., Liang, M.S., Shiramatsu, T., Wu, T.S. (1981). Electrical Characteristics of Polymer Thick Film Resistors, Part I: Experiemental Results. *IEEE Trans. on Components, Hybrids, and Manuf. Technol.*, 4(3), 283–288.

[17] Mleczko, K., Zawiślak, Z., Stadler, A.W., Kolek, A., Dziedzic, A., Cichosz, J. (2008). Evaluation of conductive-to-resistive layers interaction in thick-film resistors. *Microelectron. Rel.* 48, 881–885.

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 595–606.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

DYNAMIC SIGNAL STRENGTH MAPPING AND ANALYSIS BY MEANS OF MOBILE GEOGRAPHIC INFORMATION SYSTEM

Marcin Kulawiak, Witold Wycinka

Gdańsk University of Technology, Faculty of Electronics, Telecommunication and Informatics, G. Narutowicza 11/12, 80-233 Gdańsk, Poland (✉ Marcin.Kulawiak@eti.pg.edu.pl, +48 58 347 1728, Witek.Wycinka@wp.pl)

Abstract

Bluetooth beacons are becoming increasingly popular for various applications such as marketing or indoor navigation. However, designing a proper beacon installation requires knowledge of the possible sources of interference in the target environment. While theoretically beacon signal strength should decay linearly with log distance, on-site measurements usually reveal that noise from objects such as Wi-Fi networks operating in the vicinity significantly alters the expected signal range. The paper presents a novel mobile Geographic Information System for measurement, mapping and local as well as online storage of Bluetooth beacon signal strength in semi-real time. For the purpose of on-site geovisual analysis of the signal, the application integrates a dedicated interpolation algorithm optimized for low-power devices. The paper discusses the performance and quality of the mapping algorithms in several different test environments.

Keywords: beacon, mapping, GIS, geovisual analytics.

1. Introduction

Mapping signal strength constitutes an important issue in many challenges, related *e.g.* to indoor navigation [1]. Because of this, effective methods of mapping signal strength have been the subject of intense research for many years. Thus far, construction of signal strength maps has required the use of a complex measuring setup [2], with the collected results being processed using a dedicated desktop [3] or cloud-based software [4]. This has been primarily accomplished by the interpolation algorithms such as *Inverse Distance Weighting* (IDW) [5] and kriging [6] which have shown to produce accurate signal maps [7–10] at the cost of high computational complexity [11, 12]. Because of this, the processes of data collection, analysis and visualization have often been performed in separate hardware and software environments [13]. Moreover, on-site visualization of collected data has only been available using a laptop or a terminal device acting as a client for a cloud-based processing system [14]. However, recent developments in modern libraries dedicated to processing of geographical data have introduced the potential of collection, display and analysis of spatial data within a single instance of a *Geographic Information System* (GIS) [15, 16]. In addition, the successive advancements in the development of high-performance mobile devices has opened a pathway to the integration of data collection, storage, visualization and analysis on a single battery-powered device.

In-situ interpolation, mapping and analysis of signal strength data could quickly identify possible sources of interference and thus significantly improve *e.g.* the process of planning Bluetooth beacon placement for optimal area coverage. However, because the established data interpolation algorithms such as kriging and IDW are too computationally complex to use them on mobile devices, no such tools have been created thus far. In the above context, we propose a new mapping and analysis algorithm, optimized for use in real time on mobile devices and

Article history: received on Mar. 20, 2017; accepted on Jun. 10, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0057.

designed in accordance to the paradigms of Geovisual Analytics. The algorithm has been implemented and tested as part of a novel mobile Geographic Information System for dynamic collection, mapping and analysis of Bluetooth beacon signal strength for the purpose of indoor applications. In the paper we compare the performance and interpolation quality of the proposed algorithm with those offered by IDW and kriging for the same datasets. The interpolation results are cross-validated with in-situ measurements, and the uncertainty of produced maps is discussed. Finally, we present details of the mobile GIS architecture and discuss the results of testing the system in three distinct indoor environments.

2. Materials and methods

Recently, more and more indoor services (including *e.g.* direct marketing or navigation) are provided using electronic beacons. A beacon broadcasts its unique identifier to nearby electronic devices using a standard protocol and frequency. The most commonly used electronic beacons employ the *Bluetooth Low Energy* (BLE) standard, which has been introduced in version 4.0 of the Bluetooth specification [17]. Because BLE-compliant devices are not meant to transmit large volumes of data, the protocol has instead been optimized to provide stable connections in a range of 5 m to 15 m at very low power requirements [18]. Single-mode devices which only support BLE are commonly referred to as beacons. Beacons may be used together with any dual-mode Bluetooth 4.0 compliant device, which includes most modern smartphones [19]. Beacons are characterized by several parameters, many of which can usually be modified to some extent by the user. These include the device's transmit rate and power, its *Universally Unique Identifier* (UUID) as well as the communication protocol. Depending on the device and its configuration, Bluetooth beacons may continuously operate even for a few years [20].

The currently available beacons can be configured to operate using a variety of protocols. The most popular ones are iBeacon (developed by Apple Inc.) and Eddystone (developed by Google Inc.). The iBeacon protocol is a multi-purpose tool based on Bluetooth low energy proximity sensing. The protocol supports detection of devices which come into proximity to a single beacon, as well as monitoring the events of devices entering specific regions defined by a network of beacons [21]. However, because iBeacon is a proprietary protocol, building a universal signal analysis solution requires the use of an open protocol such as Eddystone. [22]. The Eddystone protocol is also based on the BLE standard, however it is an open specification released under the Apache 2.0 license, which makes it a cross-platform one and free to use. Currently, Eddystone supports broadcasting several types of frame data, including [23]:

- Eddystone-UID which sends out a unique identifier of the beacon, allowing for its identification by compatible devices in its proximity;
- Eddystone-URL which transmits a web address;
- Eddystone-TLM which transmits a set of the beacon’s parameters such as a battery level, temperature or humidity;
- Eddystone-EID which broadcasts a dynamically changing Beacon ID, useful for security applications.

Due to its openness and compatibility with a wide range of devices, Eddystone was selected as the protocol to be used during the presented research. As a source of signal for mapping, a selection of Bluetooth beacons implementing the Eddystone protocol was acquired. The beacons in question are shown in Fig. 1.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 595–606.

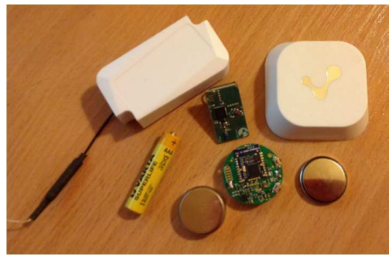

Fig. 1. Beacons used in the presented research
(from left: AprilBeacon 227A, AprilBeacon sensor 401, Kontakt.io Smart Beacon).

The detailed specifications of the beacons used in the presented research may be found in Appendix 1.

For the purpose of mobile measurement and analysis of beacon signal strength, a dedicated Geographic Information System has been designed and implemented. The system was run on a Bluetooth 4.1 compatible smartphone with 2 GB of RAM and a quad-core Snapdragon 810 processor working at a maximum frequency of 2 GHz under the control of Android 5.01 operating system. The detailed architecture of the GIS as well as the principles of its operation are presented in the following section.

2.1. System architecture and operation

Electromagnetic signals exist in a common geographical context with other physical phenomena. In computer science, physical phenomena of various nature are commonly integrated, processed and analysed with the use of Geographic Information Systems [24, 25]. Although data integration and analysis has thus far been constrained to GIS running on Desktop or Server-class computer systems, the recent advancements in both cross-platform GIS libraries and mobile computing device performance has enabled the construction of an innovative solution which integrates data collection, processing and analysis on a single low-power mobile device. The architecture of the presented solution is shown in Fig. 2.

The system is built using modern web-enabled technologies such as HTML5. This makes the system modules architecture-independent which enables their cross-platform deployment and reuse. Although the mobile application of the system is self-contained and provides whole necessary functionality directly on a mobile device, the system gives users the option of backing up their measurements on a remote server.

When installed on an Android device, the Mobile application enables the collection of signal power data from the device's Bluetooth radio. From the user's perspective, the process involves placing the device in one of the designated measurement points, which are displayed on the device's screen in the form of a grid overlaid on the map of the area. Points which have not yet been assigned any measurement values are presented on the map in red colour. The collected data are stored in the Signal buffer, where they are averaged over a user-selectable period of time (by default, the application averages 30 consecutive measurement results). The data averaged by the Signal buffer are then passed on to the Database management module. The Database management module pre-processes the collected data for storage in a database. The data may be stored in the Local database as well as server-side, where they are received by the Front controller module and passed through the Server database management module. Once the data have been stored in a database, they may be displayed by the User interface module. Both the Database management and User interface modules are implemented using the Open-Source Convertigo platform, which delivers a secured and scalable cross-platform mobile development middleware [26]. The Local database may be pre-populated with a map of the relevant area,

which may be used as reference during measurements. The map is displayed by the system's GIS module, which is also used to assign a geographical reference to the collected data. The module is built using the Open-Source OpenLayers library [27]. OpenLayers allows for the construction of interactive GIS applications for display and manipulation of vector and raster geospatial data, which can be obtained from local sources such as GML and GeoJSON files, or external ones through open protocols such as WMS and WFS. Aside from enabling the collection of measurements in a geographical context, the GIS module provides real-time interpolation and mapping of the measured signal strength with the application of Geovisual Analytics [28]. The Geovisual Analytics sub-module employs distance-weighted interpolation of the collected discrete measurement data into a continuous map which covers the entirety of the relevant area. Moreover, the Geovisual Analytics sub-module enables adaptive matching of the displayed interpolation's colour palette to the current range of signal strength. The interpolated signal strength map is presented to the end user overlaid on the map of the relevant area.

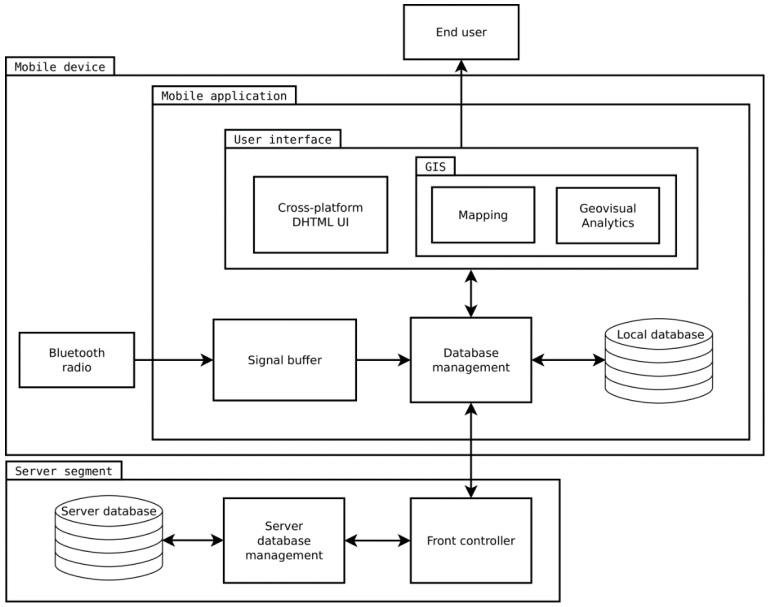

Fig. 2. Architecture of the system for mobile signal strength mapping and analysis.

2.2. Signal interpolation and mapping algorithm

Because the computational complexity of advanced interpolation algorithms such as IDW or kriging rises polynomially with the number of input points, computing high-resolution maps from a large number of measurements is not feasible on a mobile device. In response to this limitation, an optimized interpolation algorithm has been developed. The proposed algorithm assumes a regularized grid of input measurements which must be taken by the user. If a point cannot be measured (*e.g.* because of an obstacle), the value of signal at this point is interpolated using the closest known values in its horizontal and vertical neighbourhood on the grid. The final grid is then used to create a continuous map of the relevant area by interpolating values between neighbouring data points. The applied interpolation method, similarly to IDW and kriging, is derived from the Tobler Law [29] and uses distance weighting to estimate values for unmeasured locations. However, thanks to the grid generated in the first step of the algorithm,

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 595–606.

the distances are only analysed in the local instead of global context, which brings significant time savings. In particular, the value of the proposed interpolation function for a point (x, y) is given by the formula:

$$f(x,y) = \sum_{i=1}^{N} w_i(x,y) f(x_i, y_i), \tag{1}$$

where: $N \in (1; 4)$ is a number of analysed neighbouring measurement points; $f(x_i,y_i)$ is a signal value measured at point i and $w_i(x,y)$ is a weight of the i-th neighbouring measurement point, as per the uniform weight function:

$$w_i(x,y) = e^{\left(-\left(\frac{x-x_i}{dx}\right)^2 - \left(\frac{y-y_i}{dy}\right)^2\right)}. \tag{2}$$

In the above formula, a point (x_i, y_i) denotes the location of the i-th measured value and dx, dy are the x and y distances between adjacent measurement points in the grid.

The performance of the proposed interpolation algorithm has been tested on a mobile device equipped with 2 GB of RAM and a Qualcomm Snapdragon 810 quad-core CPU working with a maximum frequency of 2.0 GHz. The tests measured time required for computation and rendering of a 1080 × 1100 pixel signal map interpolated from 165 measurement points from the Large Hall dataset (which is described in detail in Subsection 3.3). The performance was averaged over ten consecutive runs. As reference, the same dataset has been interpolated on the same device using javascript IDW implementation by Manuel Bär [30]. On average, the presented algorithm rendered the complete signal map in 531 ms, which means that a new interpolation is ready in less than a second after making a new measurement. For comparison, the IDW interpolation of the same dataset took on average 7302 ms. This difference is particularly significant because IDW is known to be one of the less computationally intensive interpolation methods, especially in comparison with kriging [31, 32].

The algorithm also compares favourably with other methods when it comes to interpolation quality. Fig. 3 presents a comparison of signal maps produced from the same dataset (the Small Flat dataset, described in detail in Subsection 3.3) by the proposed method as well as by the much more computationally complex IDW and Empirical Bayesian Kriging ones.

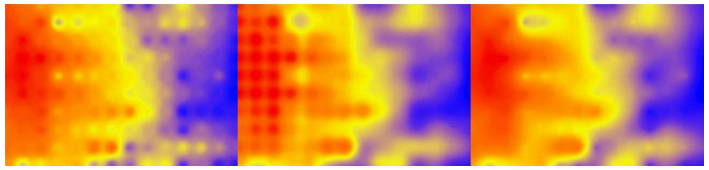

Fig. 3. Interpolations of the same set of measurements by IDW (left), the proposed method (centre) and kriging (right).

As it can be observed in Fig. 3, the proposed method delivers a smooth interpolation which retains local details, very much like kriging. IDW on the other hand overly emphasizes the individual measurements, which results in a grid-like pattern on the interpolated map. In comparison with the proposed method, kriging predicts slightly lower signal strength values near the beacon (the area to the left shown in red colour), while IDW predicts a lower signal strength value in the centre of the flat area (shown in shades of orange and yellow). At the same time, the differences between values interpolated by kriging and IDW may be found in every part of the flat area, however their average values are very similar. Computing the mean absolute deviation between the values of each map produces a relative uncertainty of 19% between the proposed method and kriging, 31% between the proposed method and IDW, and 19% between kriging and IDW.

The assessment of relative interpolation quality of the proposed method requires cross-validation of the interpolated values with in-situ measurements. For this purpose, the Small Flat dataset has been divided in two. A regular subset representing approximately 14% of all points, selected from every second row and column has been removed, and the remaining dataset has been used to perform interpolation using the proposed method as well as kriging and IDW. The smaller data subset was then used to sample the resulting rasters in the exact locations of the original measurements. Computing the mean absolute deviation between the original measurements and values predicted by each method produces a relative uncertainty of 31% for kriging, 35% for the proposed method and 50% for IDW.

According to the Bluetooth specification, a signal recorded by a Bluetooth radio has a relative value uncertainty of 6% to 10% depending on the device [17, 33]. Assuming an average relative measurement uncertainty of 0.7% and the use of a low quality Bluetooth receiver, the signal maps produced by the presented system have a relative combined standard uncertainty of $(0.19^2 + 0.35^2 + 0.1^2 + 0.007^2)^{(1/2)} = 41\%$. This is a very good result, considering the fact that if the same maps could be created on a mobile device by means of more advanced methods, their relative combined standard uncertainty would be 38% for kriging and 55% for IDW.

As it can be seen, the initial research suggests that the proposed algorithm produces signal strength maps with a quality similar to the computationally-intensive kriging, while providing a significantly better performance on a mobile device than even the less computationally complex IDW.

In the above context, the developed system has been tested in several different environments. The results of those tests are presented in the following section.

3. Results and discussion

Mapping signal strength indoors requires a reliable and appropriately powerful source. Because of this, it was necessary to test and compare characteristics of the three Bluetooth beacons in an open environment. First, every beacon was put in its maximum transmission power mode (where possible). Then, one after another, the beacons were placed on an elevated surface in an interference-free area and their signal decay characteristics were measured in one-metre distances (every recorded value was averaged from 50 consecutive measurements). Finally, the results were analysed using the application's measurement plot feature. The average relative uncertainty of the measurements was approximately 0.7%. The results may be found in Fig. 4.

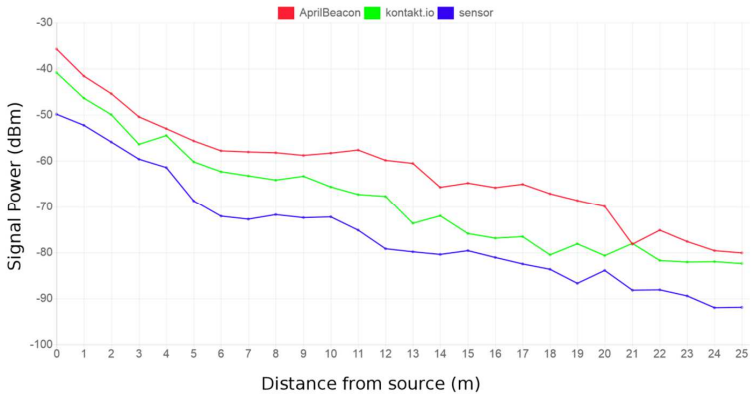

Fig. 4. Signal decay characteristics of the examined Beacons.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 595–606.

As it can be seen in Fig. 4, AprilBeacon 227A in its strongest configuration (signal power of 4 dBm) provided the best signal quality in an open environment, with the signal being readable even 25 m from the source. Despite the same power configuration (4 dBm), the signal produced by Kontakt.io was on average 6 dBm lower than AprilBeacon 227A. This may be attributed to the latter's lack of an external antenna. The AprilBeacon sensor 401 provided the lowest signal power, on average 7 dBm weaker than Kontakt.io. Basing on the above results, it was decided that AprilBeacon 227A in its strongest configuration would be used for further measurements.

Once the source of the signal was selected, the presented system has been applied to mapping and analysis of BLE signal strength in three different environments. The obtained results may be found below.

3.1. Test environment 1: Empty Hallway

The first test environment was an empty hallway with four closed wooden doors in the side walls. The walls are made of 12 cm-thick bricks. The main entrance to the hallway is through a glass door in the back. The wall opposite the main entrance has a line of glass windows. The floor is covered with lacquered rubber boards, while the suspended ceiling is made of drywall. The source beacon was placed to the right of the main entrance, approximately 1.2 m above the floor. Fig. 5 presents the recorded Beacon signal strength map for this environment (the position of the beacon is marked with x). The measurements were performed in 1 m intervals, and every recorded value was averaged from 50 consecutive measurements. The average relative uncertainty of the measurements was 0.7%.

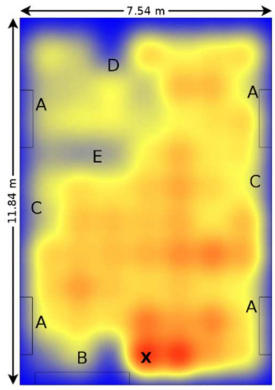

Fig. 5. A beacon signal strength map for the Empty Hallway. Red colour denotes the highest recorded signal power (–59,2 dBm), while the lowest recognizable signal (–93,3 dBm) is represented with blue colour.

The analysis of the signal strength map generated by the system shows some typical wave phenomena in the analysed environment. First of all, it is possible to observe a sharp drop in signal strength caused by the absorption of the signal through the wooden door on both sides of the hallway (points A). In those places a large part of the signal is radiated onto the outside corridors. A similar situation occurs at point B, where the signal is absorbed by the main entrance of the hallway. Interference was also likely caused by metal benches located near both side walls (points C) as well as a Wi-Fi router on the ceiling (point D). Further analysis revealed a decline in the signal strength at point E, which may be attributed to the interference caused by a Wi-Fi router on the lower floor.

3.2. Test environment 2: Large Hall

The second test environment was a large hall with walls made of 30 cm-thick breezeblocks. The hall has two sets of wooden door with glass windows on opposite side walls. There are four concrete pillars in the hall, two near each of the side walls. The floor is covered with carpeted felt. The beacon was placed near the middle of the hall approximately 1.8 m above the floor. Fig. 6 presents the recorded Beacon signal strength map for this environment (the position of the beacon is marked with x). The measurements were performed in 1 m intervals and every recorded value was averaged from 50 consecutive measurements. The average relative uncertainty of the measurements was 0.7%.

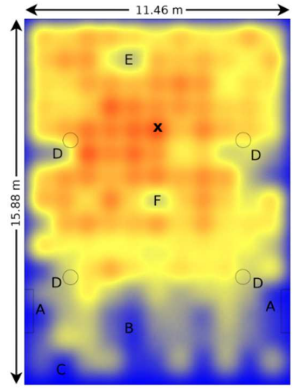

Fig. 6. A beacon signal strength map for the Large Hall. Red colour denotes the highest recorded signal power (–63,3 dBm), while the lowest recognizable signal (–93,1 dBm) is represented with blue colour.

Some of the typical phenomena observed in the previous test environment may be seen here as well. This includes signal anomalies near the wooden doors (points A) as well as interference form Wi-Fi routers inside the hall (points B and C). Aside from those, it is also possible to discern signal occlusion caused by the four concrete pillars (points D). Moreover, it is easy to spot a general loss of signal strength in the area of the concrete pillars, which is likely due to scattering caused by the round shape and smooth surface of the pillars. Further analysis attributed the decline in the signal strength in points E and F to the interference caused by Wi-Fi routers located on the upper floor.

3.3. Test environment 3: Small Flat

The final test environment was a flat in a five-storey building. The flat features load-bearing walls made of 35cm-thick concrete blocks and partition walls made of 12 cm-thick bricks. Ceilings are built of reinforced concrete with a thickness of 20 cm. The floor is covered with lacquered boards. The study area consists of a hallway, an office room with pieces of furniture such as desks, a kitchen and one room with plenty of free space. In order to minimize the impact of floor-level obstacles on the Fresnel zone, the beacon was mounted near the middle of the leftmost apartment wall at a height of 2.30 m. Fig. 7 presents the recorded Beacon signal strength map for this environment (the position of the beacon is marked with x). Because this space was considerably more tightly-packed than the previous ones, the measurements here were performed in 0.5 m intervals and every recorded value was averaged from 50 consecutive measurements. The average relative uncertainty of the measurements was approximately 0.75%.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 595–606.

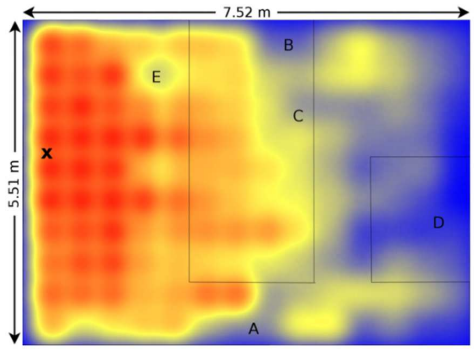

Fig. 7. A beacon signal strength map the Small Flat. Red colour denotes the highest recorded signal power (–44,3 dBm), while the lowest recognizable signal (–94,5 dBm) is represented with blue colour.

The signal map presented in Fig. 7 enables one to easily identify potential sources of interference and assess their impact on the recorded signal strength. A sharp decrease in signal power is evident in point A, where a Wi-Fi router using the 2.4 GHz band was placed. A similar situation may be observed in point B, where a Wi-Fi client computer was located. Moreover, signal-shielding metal objects (such as a large wall-mounted heater) were situated in the vicinity of measurement point C. Similarly, kitchen appliances such as an oven and a dishwasher are responsible for noticeable decreases in the recorded signal power for point D. Further analysis revealed that the sharp drop in the signal power in point E was caused by a Wi-Fi router located on the floor below. Analysis of the signal strength map of the entire apartment reveals a significant decline in strength of the signal passing through the brick walls of the room in the middle. Moreover, a decrease in the recorded signal power was even greater behind walls covered with ceramic tiles, such as those in the kitchen (around point D).

4. Summary

The paper presents the development and first field tests of an innovative mobile Geographic Information System for Bluetooth beacon signal mapping. Testing the system in several different environments has proved that it is capable of providing on-site signal strength interpolation and mapping in real time on a plain smartphone with Android OS. Moreover, the applied dedicated interpolation algorithm has shown to produce relatively high quality results, comparable to those offered by kriging, and better than those provided by IDW (both of them being much more computationally complex interpolation methods). This being said, it should be noted that the applied mobile optimizations require the algorithm to work with a regularized measurement grid and thus should not be applied to the interpolation of sparse irregular measurements (which in turn are known to be handled properly by both kriging and IDW). Still, the applied optimizations enable the system to work in semi-real time, producing signal strength maps directly on a mobile device during measurements. Moreover, the system applies Geovisual Analytics for adaptive signal strength colour palette matching, which ensures good visibility of local signal power spikes and enables easy identification of possible sources of interference. The interpolated signal strength map is presented to the end user overlaid on the map of the relevant area.

The presented results indicate that features like real-time signal measurement, interpolation and analytical mapping, coupled with cloud backup capabilities, make the system a flexible tool for researchers as well as engineers working on indoor beacon applications such as proximity sensing and navigation.

References

[1] Liu, S., Chen, Y., Trappe, W., Greenstein, L.J. (2009). Non-interactive localization of cognitive radios based on dynamic signal strength mapping. *2009 Sixth International Conference on Wireless On-Demand Network Systems and Services*, Snowbird, UT, 85–92.

[2] Yin, J., Yang, Q., Ni, L.M. (2008). Learning adaptive temporal radio maps for signal-strength-based location estimation. *IEEE Transactions on Mobile Computing*, 7(7), 869–883.

[3] Ji, Y., Biaz, S., Pandey, S., Agrawal, P. (2006). ARIADNE: a dynamic indoor signal map construction and localization system. *Proc. of the 4th international conference on Mobile systems, applications and services,* 151–164.

[4] de Moraes, L.F.M., Nunes, B.A.A. (2006). Calibration-free WLAN location system based on dynamic mapping of signal strength. *Proc. of the 4th ACM international workshop on Mobility management and wireless access*, 92–99.

[5] Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. *Proc. of the 1968 23rd ACM national conference*, 517–524.

[6] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8), 1246–1266.

[7] Connelly, K., Liu, Y., Bulwinkle, D., Miller, A., Bobbitt, I. (2005). A toolkit for automatically constructing outdoor radio maps. *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, 248–253.

[8] Phillips, C., Ton, M., Sicker, D., Grunwald, D. (2012). Practical radio environment mapping with geostatistics. *2012 IEEE International Symposium on Dynamic Spectrum Access Networks*, Bellevue, WA, 422–433.

[9] Wielgosz, P., Grejner-Brzezinska, D., Kashani, I. (2003). Regional ionosphere mapping with kriging and multiquadric methods. *Journal of Global Positioning Systems*, 1(4), 48–55.

[10] Lee, H. K., Li, B., Rizos, C. (2005). Implementation procedure of wireless signal map matching for location-based services. *Proc. of the Fifth IEEE International Symposium on Signal Processing and Information Technology*, 429–434.

[11] Kerry, K.E., Hawick, K.A. (1998). Kriging interpolation on high-performance computers. *International Conference on High-Performance Computing and Networking*, 429–438.

[12] Murphy, R.R., Curriero, F.C., Ball, W.P. (2009). Comparison of spatial interpolation methods for water quality evaluation in the Chesapeake Bay. *Journal of Environmental Engineering*, 136(2), 160–171.

[13] Ye, S.J., Zhu, D.H., Yao, X.C., Zhang, X., Li, L. (2016). Developing a mobile GIS-based component to collect field data. *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, Tianjin, 1–6.

[14] Han, W., Hu, Y., Zhang, J., Liu, Q. (2015). Mobile Data Acquisition and Management System Design Based on GIS and GPRS. *Metallurgical and Mining Industry*, 2, 243–249.

[15] Moszynski, M., Kulawiak, M., Chybicki, A., Bruniecki, K., Bieliński, T., Łubniewski, Z., Stepnowski, A. (2015). Innovative Web-Based Geographic Information System for Municipal Areas and Coastal Zone Security and Threat Monitoring Using EO Satellite Data. *Marine Geodesy*, 38(3), 203–224.

[16] Kulawiak, M., Kulawiak, M. (2017). Application of Web-GIS for Dissemination and 3D Visualization of Large-Volume LiDAR Data. *The Rise of Big Spatial Data*, Springer International Publishing, 1–12.

[17] Bluetooth Special Interest Group. Specification of the Bluetooth® System, version 4.2. 2014. *https://www.bluetooth.org/DocMan/handlers/DownloadDoc.ashx?doc_id=286439* (Nov. 2016).

[18] Gomez, C., Oller, J., Paradells, J. (2012). Overview and evaluation of Bluetooth low energy: An emerging low-power wireless technology. *Sensors*, 12(9), 11734–11753.

[19] Townsend, K., Cufí, C., Davidson, R. (2014). *Getting started with Bluetooth low energy: tools and techniques for low-power networking*. O'Reilly Media, Inc.

[20] Mackensen, E., Lai, M., Wendt, T.M. (2012). Performance analysis of an Bluetooth Low Energy sensor system. *2012 IEEE 1st International Symposium on Wireless Systems (IDAACS-SWS)*, Offenburg, 62–66.

[21] Getting Started with iBeacon. (2014). Apple Inc. *https://developer.apple.com/ibeacon/Getting-Started-with-iBeacon.pdf* (Nov. 2016).

[22] iBeacon – Frequently Asked Questions. (2014). Cisco Inc. *http://www.cisco.com/c/dam/en/us/solutions/collateral/enterprise-networks/connected-mobile-experiences/ibeacon_faq.pdf* (Nov. 2016).

[23] Eddystone protocol specification. (2016). *https://github.com/google/eddystone/blob/master/protocol-specification.md* (Nov. 2016).

[24] Moszynski, M., Chybicki, A., Kulawiak, M., Lubniewski, Z. (2013). A novel method for archiving multibeam sonar data with emphasis on efficient record size reduction and storage. *Polish Maritime Research*, 20(1), 77–86.

[25] Kulawiak, M. (2016). Operational algae bloom detection in the Baltic Sea using GIS and AVHRR data. *Baltica*, 29(1), 3–18.

[26] *www.convertigo.com* (Apr. 2017).

[27] *openlayers.org* (Apr. 2017).

[28] Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M. J., MacEachren, A., Wrobel, S. (2007). Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, 21(8), 839–857.

[29] Tobler, W. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2), pp. 234–240.

[30] Bär, M. *www.geonet.ch* (Apr. 2017).

[31] Reed, P.M., Ellsworth, T.R., Minsker, B.S. (2004). Spatial interpolation methods for nonstationary plume data. *Ground Water*, 42(2), 190–202.

[32] Murphy, R.R., Curriero, F.C., Ball, W.P. (2009). Comparison of spatial interpolation methods for water quality evaluation in the Chesapeake Bay. *Journal of Environmental Engineering*, 136(2), 160–171.

[33] Bluetooth Special Interest Group. Specification of the Bluetooth® System, version 1.0. 1999. *http://ece.wpi.edu/analog/resources/bluetooth_a.pdf* (Nov. 2016).

Appendix 1: Detailed specifications of used Bluetooth beacons

Table 1. Specifications of Kontakt.io Smart Beacon.

Parameter	Value
Transmission Power	–30 dBm to 4 dBm
Sensitivity	–93 dBm
Working temperature	–20ºC to +60ºC
Processor	32-bit Arm Cortex MO CPU core
Bluetooth processor	Nordic nRF51822
Data rate	250 kb/s to 2 Mb/s
Flash memory size	256 kB
Ram size	16 kB
Battery	1x 1000 mAh CR2477
Exchangeable battery	yes
Battery life at 350 ms broadcast interval	2 years
External antenna	no
Minimum number of units sold	3
Price per unit	27 $

Table 2. Specifications of AprilBeacon 227A.

Parameter	Value
Transmission Power	–23 dBm to 4 dBm
Working temperature	–40ºC to +85ºC
Processor	Texas instruments 8051
Bluetooth processor	Texas instruments CC2540
Data rate	250 kb/s to 2 Mb/s
Flash memory size	256 kB
Battery	2x AAA (1000 mAh)
Exchangeable battery	yes
Battery life at 100 ms broadcast interval	4,5 month
External antenna	yes, 50 Ω
Price per unit	12 $

Table 3. Specifications of AprilBeacon sensor 401.

Parameter	Value
Working temperature	–40ºC to +85ºC
Processor	Texas instruments 8051
Bluetooth processor	Texas instruments CC2541
Data rate	250 kb/s to 2 Mb/s
Flash memory size	256 kB
Ram size	8 kB
Battery	1x 620 mAh CR2450
Exchangeable battery	yes
Battery life at 100 ms broadcast interval	2 months
Sensors	Light sensor, accelerometer, vibration sensor
External antenna	no
Price per unit	22 $

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 607–616.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

LOW HUMIDITY CHARACTERISTICS OF POLYMER-BASED CAPACITIVE HUMIDITY SENSORS

Jacek Majewski

Lublin University of Technology, Faculty of Electrical Engineering and Computer Science, Nadbystrzycka 38A, 20-618 Lublin, Poland
(✉ j.majewski@pollub.pl, +48 81 538 4314)

Abstract

Polymer-based capacitive humidity sensors emerged around 40 years ago; nevertheless, they currently constitute large part of sensors' market within a range of medium (climatic and industrial) humidity 20–80%RH due to their linearity, stability and cost-effectiveness. However, for low humidity values (0–20%RH) that type of sensor exhibits increasingly nonlinear characteristics with decreasing of humidity values. This paper presents the results of some experimental trials of CMOS polymer-based capacitive humidity sensors, as well as of modelling the behaviour of that type of sensor. A logarithmic functional relationship between the relative humidity and the change of sensor output value at low humidity is suggested.

Keywords: polymer-based capacitive humidity sensors, low humidity measurement, humidity sensors modelling.

1. Introduction

The modern humidity measurements in science and industry emerged around two centuries after the development of early instruments for temperature measurement. De Saussure (1783) built the first hair-tension hygrometer, based on the interaction of water molecules with keratin of a grease-free hair which has a polymeric structure. During the 19th century, a psychrometer was invented in 1825 (August) and later enhanced into an aspiration psychrometer (by Assmann, 1887, for high altitude balloons). Simultaneously, dew-point hygrometers were introduced (Daniell, 1820; Lambrecht, 1881) [1]. In the 20th century a lithium chloride-heated resistive sensor was proposed by Dunmore in 1938 (and patented in 1942) [2]. This sensor output an electrical signal and offered a shorter response time than former hygrometers; moreover, it was feasible to use it in constructing remote indication systems like radiosondes for meteorology.

In 1973, the world's first miniaturised thin-film polymer-based capacitive humidity sensor, trade-marked HUMICAP®, was introduced by Vaisala Oy company in two types: for radiosondes, and for general purpose use (*e.g.* in hygrometers, control systems) [3]. So, starting with humidity-dependent mechanical properties of the polymeric keratin structure of hair, hygrometry turned full circle back to polymeric humidity-sensing materials, due to their dielectric properties. The main competitor of the polymer-based design among humidity sensors is the aluminium oxide sensor [4]; also a thin-film porous structure, but prone to calibration drift and a low response rate, more expensive and more delicate than polymer-based sensors. The Al_2O_3-based sensors' response signal is proportional to the absolute rather than relative humidity.

A thin (ca. 1 μm) polymer film is advantageous, because the capacitance of a parallel-plate capacitor increases inversely proportionally to the film thickness. A thinner film could have

Article history: received on Jun. 06, 2016; accepted on Jul. 31, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0048.

a less homogeneous structure. Even more important is the reduced response time of thin-film sensors for step changes of relative humidity (of the order of seconds).

Since the sensor is a kind of capacitor, the thin film of polymer is sandwiched between two or three metallic electrodes, usually planar (devices with cylindrical geometry are tested as beneficial to obtain a shorter response time). The trouble with a two-electrode sandwich is that the upper electrode must be porous enough to allow water molecules to penetrate freely into the polymer layer, and at the same time the electrical continuity and imperviousness to non-water molecules must be secured. So, the optimum thickness for used metals (gold, chromium, nickel) is ca. 10 nm [5]. Connection of such an ultra-thin metal layer to an electric tap is a difficult technological operation. To avoid this, in many arrangements the upper porous thin electrode is not connected, since as a zero-potential one it ensures parallel running of the electric field lines through the polymer. The two electrically contacted, interdigitated bottom electrodes are placed on a thick and stiff glass substrate.

In some devices, the upper electrode is thick but comb-shaped, to allow better, rapid penetration of water molecules into polymer, although then the active surface of capacitor is reduced by half. In another design only two bottom interdigitated electrodes are applied without the upper porous electrode, and the polymer film is deposited in the last stage of fabrication process; however, the lines of electric field are curved and not parallel.

Many polymeric materials have been tested as humidity sensitive layers, the main feature of which is the presence of a so called free volume, estimated at around 30% of the total volume of the layer. The free volume is a network composed of pores, micro-voids (cavities) and micro-channels, interconnected and characterised by statistical distributions. The polymers for humidity capacitive sensors should be thermally stable and chemically resistive, and the polymer relative permittivity should be low (within a range from 3 to 10 [6]). The detailed composition of a polymer is usually a top secret of its manufacturer.

More than 70% of all humidity sensors are the polymer-based capacitive sensors [7], because they offer a very broad range of quasi-linear characteristic of change in output signal versus relative humidity (usually 20–80% RH, and 10–90% RH in improved designs). In a range of 90–100% RH the long-term stability of the sensors becomes poor, hysteresis large, and permanent offset to the sensor can remain. On the side of low humidity: 0–10 (20)% RH, the static characteristic's linearity falls off, and the response time becomes longer; on the other hand, hysteresis is negligible.

In many applications, like generation of pure materials (*e.g.* gases [8]), detection of a trace moisture content in natural gas pipelines, drying of solid materials, in meteorology [9] at high altitudes (climate change studies), or in cosmonautical observations [10, 11], the measurement of low humidity is essential. Some companies meet that need with sensors dedicated to low humidity (*e.g.* HM-1520 Humirel, DRYCAP® DMT 150 Vaisala Oy, K5-W IST AG, HC103M2 E+E Elektronik Ges.m.b.H). For better understanding of different behaviour of the polymer-based capacitive sensors at low and medium humidities, a model of their characteristics at low humidity would be useful.

In this paper, some novel measurement results are presented, and an attempt to suggest possible explanation of the behaviour of the polymer-based capacitive sensors at low humidity is made. Many manufacturers of such sensors often report only a drop in sensor's accuracy in a range of low humidity (the value of maximum error). If a relationship between the low RH values provided by the sensor and the reference values was established, it would help to perform measurements in this range with better accuracy.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 607–616.

2. Modelling of polymer-based capacitive humidity sensors

First of all, it should be reminded that no "air saturation with water vapour" takes place. The "saturation" process in a "vacuum – liquid water" system is profoundly the same as in an "air – liquid water" one, and ruled by the Boltzmann's distribution of energy of water molecules on the surface of liquid phase. The amount of water vapour depends on the number of water molecules that reached the "escape energy" limit at the right tail of distribution, which in turn depends on the absolute temperature. The term "saturation" means here the state of dynamic equilibrium between liquid water and water vapour.

When in 1980's extensive research on polymer materials for humidity sensors has begun, the question of the theory of polymer-based capacitive humidity sensors operation was posed. If the relative humidity is defined as a ratio of the partial pressure of water vapour p_{wv} and the partial pressure of the saturated water vapour p_s at the same temperature, then how can the polymer-based sensors respond to that relative, instead of absolute humidity (as Al_2O_3-based sensors do respond)?

In 1985, Denton *et al*. [12] suggested that the water molecules follow the Fickian diffusion, and that the molecules inside pores in the polymer layer are in the vapour phase. However, that model did not explain why the number of water molecules should be proportional to the relative humidity. Also, the number density of water molecules in vapour (or specific humidity) is much less than the number density of water molecules in liquid water inside pores.

In 1995 Anderson [13] proposed an alternative model of operation of the polymer-based capacitive sensors. Since water molecules are very small (ca. 0.2 nm) and highly polar, all solid surfaces in contact with air are coated with a layer of physio-sorbed water molecules, attracted mostly by the van der Waals' forces. There are different polar sites in polymers at which the water molecules can be bound in various ways, *e.g.* between adjacent polymer chains. That means that the inner surfaces of free volume network, interconnected inside a polymer film, should be covered with one or more layers of water molecules. In the Anderson's model, the first layer adheres closely to the polymer inner surfaces because of hydrogen bonds (relatively strong), whereas next layers – if present – are bound with forces exponentially weaker (mainly the van der Waals' forces). The volume of voids inside the polymer is filled with water vapour under a partial pressure equal to the partial pressure in the ambient air (see Fig. 1).

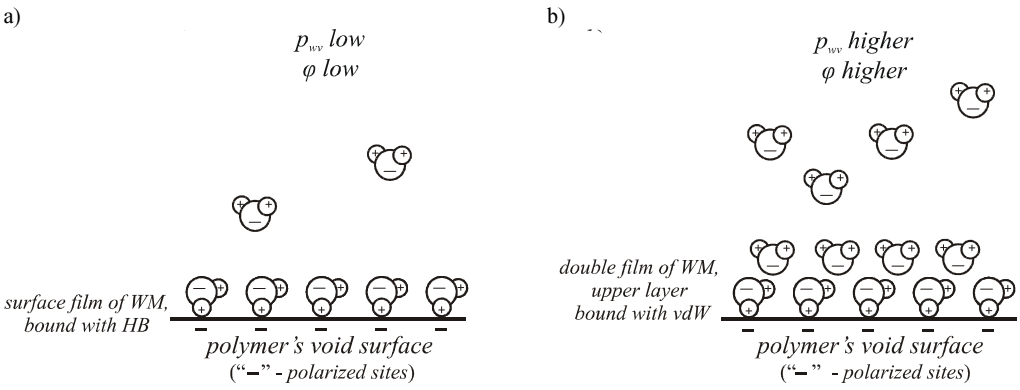

Fig. 1. A schematic illustration of the idea of change in relative humidity φ (*i.e.* under the water vapour partial pressure p_{wv}) within voids in polymeric materials; a thin *WM* film case (a); a thicker *WM* film case (b) (*HB* – hydrogen bond, *vdW* – the van der Waals' forces, *WM* – water molecules).

The polymers used for sensing layers in modern capacitive relative humidity sensors are mainly polyimides (although Vaisala Oy used cellulose acetate). Polyimides are heterocyclic polymers which offer some very attractive features when used in capacitive sensors: outstanding thermal-oxidative and chemical stability, high glass transition temperatures, high radiative- and solvent-resistance; they exhibit very good dielectric and mechanical properties with good processability; the dielectric constant relative to water is low [14, 15]. They are also easily integrated into a CMOS type integrating process. The adsorption sites of polyimide chains are mainly oxygen atoms of carboxyl groups C = O, and to some extent also nitrogen atoms in N–C groups; some structures of the polyimide family polymers contain also ether groups C–O–C but their oxygen atoms are only slightly involved in the H-bonding process. Oxygen and nitrogen atoms are strongly electronegative; hence, the adsorption sites on the polyimide backbone attract hydrogen atoms of water molecules and enable adsorption of these molecules based on hydrogen bonding.

In examination of the process of water sorption and uptake in polymers, many sophisticated measurement techniques were applied, mainly gravimetric and vibrational spectrometric ones (*e.g.* NMR, 2D-FTIR, ATR); even experiments with the use of molecules of D_2O (heavy water) and T_2O (super-heavy water) were performed. Extensive research of vibrational spectra of humidified polyimide has revealed that on the surface of thin polymer layer water molecules are mostly physio-sorbed by the hydrogen bonding interaction (whereas on the surface of some oxides, *e.g.* Al_2O_3, they undergo chemisorption which causes dissociation of these molecules and creation of surface hydroxyl groups [16]). Diffusion in polyimides depends on their crystallinity and density, as well as on chain stiffness [17]. In FTIR spectra shifts were observed which can be attributed to carbonyl groups as proton acceptors in dipole-dipole interactions with water molecules, which exhibit a strong intrinsic permanent electric dipole moment [18]; these shifts were fully reversible during the water desorption process. In [19] the activation energy of bonds inside polymer was estimated to be of the order of 1 kJ/mol which is closer to the van der Waals' bonds, and may be attributed to the bonds between water molecules in micro-pores inside the polyimide layer. That makes some authors formulate a hypothesis of two different water species present inside polyimide: one of single molecules hydrogen-bonded to polymer adsorption sites – which undergo faster sorption - and the other aggregated by the van der Waals' forces in water clusters (or H-bonded in dimers) [18]. In that research, the authors also claim that only one third of all imide interaction sites is available for water molecules; the rest is being involved in intermolecular charge transfer interactions.

In a range of medium and high humidity, in a multilayer shell of water molecules, the physio-sorption based on dipole-dipole or dipole-induced dipole could take place. Intermolecular dispersion forces are unlikely to take part because of strong polar nature of water molecules (no instantaneous dipoles). Ion-dipole interactions are also unlikely to occur because of good dielectric properties of the used polymer layer (no free ions, or only a negligible amount of them). On the other hand, in a range of low humidity, the one-molecule shell of water could cover the polymer surfaces, and H-bonding would dominate. In fact, most sorption measurements start from 10% RH upwards, and the signals of vibrational spectra for lower humidity are too weak.

The water vapour inside pores should be in equilibrium with the upper layer of water molecules' film on the surfaces of free volume network. For example, if the relative humidity outside the sensor increases, more water molecules diffuse into the polymer, and the thickness of the inner water film coating the inner surfaces increases till the increased water vapour pressure over that thicker film reaches a new equilibrium with the ambient partial pressure of water vapour. That Anderson's model accounts also for a weak dependence of the sensor's response on temperature, despite of a strong increase of the saturation partial pressure of water vapour with temperature.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 607–616.

The idea of that model was generally respected (*e.g.* [20]), although its accuracy was rather rough, and in the Anderson's equations both the polymer and water permittivity are not explicitly included. For that reason, many researchers applied the equations based on the modified Classius-Mossotti equation [21] defined as:

$$\frac{\Delta N\alpha}{3\varepsilon_0}=\frac{\varepsilon_r(\varphi)-1}{\varepsilon_r(\varphi)+2}-\frac{\varepsilon_r(0)-1}{\varepsilon_r(0)+2}, \tag{1}$$

where: ΔN is an increase in the number of dipoles (water molecules) per unit volume in the polymer film (a number density, in $1/m^3$) due to the increase in relative humidity from 0 to φ; α is a molecular polarizability (in Cm^2/V); ε_0 is the vacuum permittivity (in F/m); $\varepsilon_r(\varphi)$ is a relative permittivity of the polymer film at a given relative humidity φ; and $\varepsilon_r(0)$ is a relative permittivity of the dry film. Practically, if $\varepsilon_r(0) = 3$, then at φ = 100 %RH, $\varepsilon_r(\varphi) = 3.9$; the change in relative permittivity $\Delta\varepsilon_r = \varepsilon_r(\varphi) - \varepsilon_r(0)$ caused by uptake of water molecules is usually small, although the relative permittivity of (highly polarised) water is around 80.

When the influence of temperature on the relative permittivity is taken into account, the modified Debye equation [22, 23] is applied:

$$\frac{\varepsilon_r(\varphi)-1}{\varepsilon_r(\varphi)+2}=\frac{N}{3\varepsilon_0}\left(\alpha+\frac{\mu^2}{kT}\right), \tag{2}$$

where: μ is a dipole moment of one water molecule (in SI units: C·m); k is the Boltzmann's constant, and T is the absolute temperature (in K).

Even more precise is the modified Kirkwood's equation [24] for a binary system of dielectric materials, when the number density of polymer is practically independent of the humidity:

$$\frac{(\varepsilon_r(\varphi)-1)(2\varepsilon_r(\varphi)+1)}{9\varepsilon_r(\varphi)}=\frac{(\varepsilon_r(0)-1)(2\varepsilon_r(0)+1)}{9\varepsilon_r(0)}+\frac{4N}{3\varepsilon_0}\left(\alpha+g\frac{\mu^2}{3kT}\right), \tag{3}$$

where g is the Kirkwood correlation factor, a measure of local ordering of the dipoles: if fixing a position of one dipole does not disturb the remaining positions of the neighbouring dipoles, then g = 1. Generally, g can be a function of water molecules' uptake. Another feature is that water in confined systems behaves differently from bulk liquid water; the relative permittivity depends on the average size of volume where water is confined [25, 26].

Instead of the theoretically derived formulae containing $\varepsilon_r(\varphi)$, some researchers base on the empirical equation by Looyenga for a mixture of two dielectric materials [27]:

$$\varepsilon_r(\varphi)=\left[\gamma\left(\sqrt[3]{\varepsilon_w}-\sqrt[3]{\varepsilon_r(0)}\right)+\sqrt[3]{\varepsilon_r(0)}\right]^3, \tag{4}$$

where γ is a volume fraction of water absorbed in the polymer, and ε_w is a relative permittivity of water, which can be calculated from the following formula [28]:

$$\varepsilon_w=78.54\cdot\left[1-4.6e^{-4}(T-T_0)+8.8e^{-6}(T-T_0)^2\right], \tag{5}$$

where T_0 = 298 K.

In the above mentioned research, the aim was to evaluate dynamic changes in time within the sensor polymer's volume by simulation. For that purpose, the water concentration was calculated from the Fickian diffusion equation, and the relative humidity could be obtained using the Henry's law. However, in these formulae the relative humidity p_{wv}/p_s is not explicitly included. There is a need for a mathematical model taking into account both relative humidity and dielectric constants, also for the low humidity range; that model would be helpful

in estimation of metrological properties of polymer-based capacitive sensors applied to measurements of low humidity values.

3. Experimental setup for measurement of low humidity characteristics

In order to establish the nonlinearity of polymer-based capacitive humidity sensors in the low humidity region, an experimental setup shown in Fig. 2 was used.

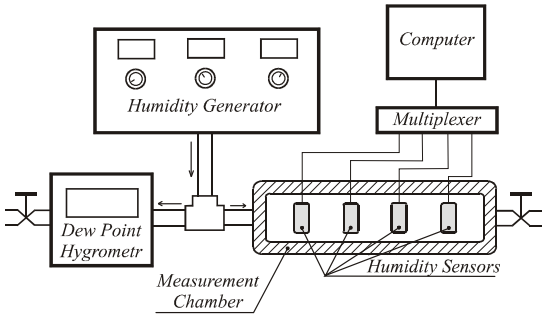

Fig. 2. A schematic of the experimental stand for measuring low humidity characteristics of humidity sensors.

The source of low humidity was a specialized trace humidity generator DG-4 (Michell Instruments/UK). An adjusted low humidity value was obtained by mixing dry air with minute amounts of humid air at a controlled flow rate. The reference instrument, a dew/frost point hygrometer GE Optica 1311 XR (General Eastern Sensing/USA) can measure trace humidity down to −80ºC frost point; its accuracy was confirmed with an NPL-traceable certificate of calibration. A set of four humidity sensors SHT 21 (Sensirion/Switzerland) was placed in a special thick-walled (for temperature equalizing) measurement chamber made of stainless steel with small orifices for mounting the sensors. The humidity generator was connected with both the dew point hygrometer and the measurement chamber by stainless steel tubing, with electro-polished inner surfaces for reducing the risk of condensation. The air flow with a precisely adjusted low humidity level was supplied from the trace humidity generator and divided in a T-shape fitting with a gas flow flux of ca. 200 l/h. In this experiment, the frost point temperature values were set within an interval from –40ºC (0.45% RH) to –10ºC (9.25% RH), stepwise with a step of approximately 5 K. The ambient temperature and pressure were controlled during the experiment. The temperature during measurements was (24.0 ± 0.2)ºC; the research lab was air-conditioned.

The sensors under test were SHT-series sensors manufactured by the market-leading Swiss company Sensirion [29]. This choice was motivated by the fact that the SHT sensors are made in CMOS technology which enables to ensure high accuracy (within a 20 to 80% RH range, 3% RH for SHT-71 and 2% RH for SHT-21, and 1.8% RH for SHT-75 within a 10–90% RHT range). These sensors are highly integrated devices; each sensor is individually calibrated and tested, and an electronic identification code is stored on the sensor chip. Each chip contains also a band-gap temperature sensor, an amplifier, an A/D converter, a programmable memory, and a digital interface. For data logging, the manufacturer offers also an evaluation kit EK-H4 (multiplexing device) with hardware and software to interface the sensors with a computer. In the reported sensor trial measurements four SHT-21 sensors, four SHT-71 sensors, and one SHT-75 sensor were tested. Every measurement point was determined after at least 1 hour of keeping a constant humidity level to assure the equilibration of water vapour concentration in the whole system.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 607–616.

4. Results and discussion

As a result of the experimental trials, three sets of low-humidity sensor characteristics were obtained. In Fig. 3 plots of the differences $\Delta\varphi$ between the relative humidity value φ_i provided by *i*-th SHT-21 sensor and the reference value φ_R measured with the dew/frost point hygrometer are shown. For each sensor, six measurement points were determined, starting from the –40ºC frost point temperature.

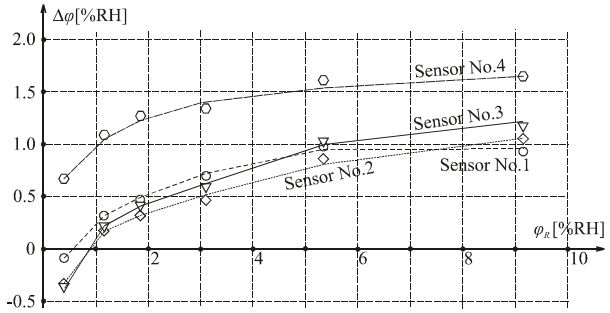

Fig. 3. Plots of the differences $\Delta\varphi$ between the relative humidity value φ_i provided by *i*-th SHT-21 sensor and the reference value φ_R.

It can be seen that $\Delta\varphi$ gradually falls off from linearity with decreasing the frost point temperature (and the relative humidity), and that for sensors Nos1–3 the plots practically overlap, whereas for sensor No. 4 the shape of plot is very similar, albeit shifted up by an offset value. The highest correlation coefficients were obtained for fitting these characteristics with a logarithmic approximation function.

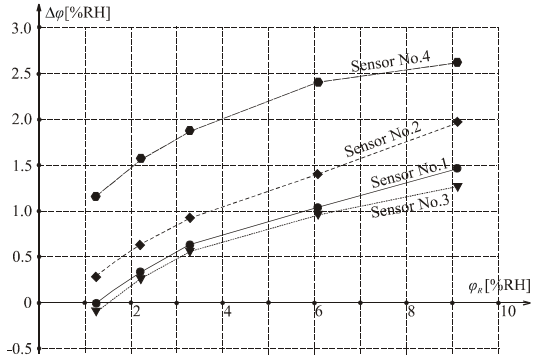

Fig. 4. Plots of the differences $\Delta\varphi$ between the relative humidity value φ_i provided by *i*-th SHT-71 sensor and the reference value φ_R.

In Fig. 4 plots of the differences $\Delta\varphi$ between the relative humidity value φ_i provided by *i*-th SHT-71 sensor and the reference value φ_R measured with the dew/frost point hygrometer are shown. For each sensor, five measurement points were determined, starting from the lowest value –30ºC of frost point temperature – because it turned out that for humidity below 0.5% RH, the SHT-71 and SHT-75 sensors provided a dummy value of 0.1% RH. Actually, only above 0.5%RH the displayed measurement data were calculated from raw processed integer codes by the sensors' electronics. Despite of the operating range declared in the technical datasheet from

0%RH to 100%RH, the interval 0–0.5% RH is excluded from the sensors' scope. In the SHT-21 sensors (newer than SHT-71) such a limitation was not noticed. The plots are similar to the SHT-21 ones, although look slightly skewer. Also for SHT-71, the highest correlation coefficients were obtained for fitting these characteristics with a logarithmic approximation function.

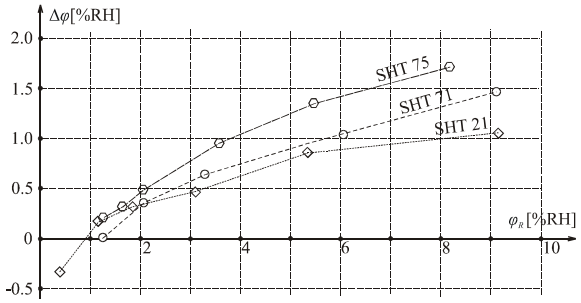

Fig. 5. Comparison of plots of the differences $\Delta\varphi$ between the relative humidity value φ_i provided by SHT-21, SHT-71 and SHT-75 sensor items, and the reference value φ_R.

The plots of the differences $\Delta\varphi$ between the relative humidity value φ_i provided by SHT-21, SHT-71 and SHT-75 sensor items and the reference value φ_R are shown in Fig. 5. The characteristic of SHT-75 exhibits similarity to SHT-71 rather than to SHT-21 one. It seems that the better accuracy declared for SHT-75 for medium humidity values has no impact on the accuracy within the low humidity range.

5. Conclusions

The tested polymer-based capacitive humidity sensors manufactured in CMOS technology exhibit in a range of low relative humidity 0–10% RH a nonlinearity of characteristics which increases when the humidity decreases towards 0% RH. The shapes of plots of these characteristics are similar, and it seems that a kind of regularity, independent of sensor's chip structure, takes place. The best fit approximation function of this nonlinearity shows a logarithmic dependency. Because some models of polymer sorption equilibrium (*e.g.* Flory-Huggins [24]) contain logarithmic dependencies, it could be possible that some interaction processes in water vapour molecules-polymer chains inside polymeric materials could be described by logarithmic relationships which could not be explicitly observed within a range of medium humidity values, but could only be revealed at low humidity (that might be *e.g.* hydrogen bond interactions of logarithmically distributed strengths.) A model of the behaviour of polymer-based capacitive humidity sensors for low humidity values would be useful to explain that nonlinearity, which suggests a logarithmic relationship between the relative humidity and the output value of the sensor.

Acknowledgements

The author is indebted to the team of the Working Group 8.1 at the Bundesanstalt für Materialforschung und -prüfung (Berlin, Germany) for practical support and scientific discussion.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 607–616.

References

[1] Rübner, K., Balköse, D., Robens, E. (2008). Methods of humidity determination. Part I: Hygrometry. *J. Thermal Anal. Cal.*, 94(3), 669–673.

[2] Dunmore, F. (1938). An Electric Hygrometer and Its Application to Radio Meteorography. *J. Res. Natl. Bur. Stand.*, 20,723–744.

[3] Salasmaa, E. (1986). Humicap® thin film humidity sensor. Gensler, W.G.(ed.). *Advanced Agricultural Instrumentation*. Dordrecht: Martinus Nijhoff Publishers.

[4] Jason, C.A., Lees A. (1964). Hygrometric elements. US Patent 3,121,853.

[5] Farahani, H., Wagiran, R., Hamidon, M.N. (2014). Humidity Sensors Principle, Mechanism and Fabrication Technologies: A Comprehensive Review. *Sensors*, 14, 7881–7939.

[6] Hársanyi, G. (1995). *Polymer Films in Sensor Applications*. Basel: Technomic Publishing AG.

[7] Rittersma, Z.M. (2002). Recent achievments in miniaturised humidity sensors – a review of transduction techniques. *Sens. Actuators A*, 96,196–210.

[8] Hübert, T., Tiebe, C., Detjens, M., Majewski, J. (2016) On-Site Calibration System for Trace Humidity Sensors. *Measurement*. In Press, Accepted Manuscript 2016.05.13, doi:http://dx.doi.org/10.2016/j.measurement.

[9] Story, P.R., Galipeau, D.W., Mileham, R.D. (1995). A study of low-cost sensors for measuring low relative humidity. *Sens. Actuators B*, 24–25,681–685.

[10] Harri, A.M., *et al.* (2014). Mars Science Laboratory relative humidity observations: Initial results. *J. Geophys. Res. Planets*, 119, 2132–2147.

[11] Rübner, K., Balköse, D., Robens, E. (2008). Methods of humidity determination. Part II: Determination of material humidity. *J. Thermal Anal. Cal.*, 94(3), 675–682.

[12] Denton, D.D., Camou, J.B., Senturia, S.D. (1985). Effects of moisture uptake on the dielectric permittivity of polyimide films. *Int. Symp. on Moisture and Humidity*, Research Triangle Park, N. C., Instrument Soc. Amer., 505–513.

[13] Anderson, P.S. (1995). Mechanism for the Behavior of Hydroactive Materials Used in Humidity Sensors. *J. Atmos. and Oceanic Technol.*, 12, 662–667.

[14] Yang, D.K., Koros, W.J., Hopfenberg, H.B., Stannett, V.T. (1985). Sorption and Transport Studies of Water in Kapton*Polyimide.I. *J. Appl. Polymer Sci.*, 30, 1035–1047.

[15] Musto, P., Mentisieri, G., Lavorgna, M., Scarinzi, G., Scherillo, G. (2012). Combining Gravimetric and Vibrational Spectroscopy Measurements to Quantify First- and Second-Shell Hydration Layers in Polyimides with Different Molecular Architectures. *J. Phys. Chem. B*, 116, 1209–1220.

[16] Korotcenkov, G. (2013). *Handbook of Gas Sensor Materials: Properties, Advantages and Shortcomings for Applications Volume 1: Conventional Approaches*. Springer Science+Business Media, LLC.

[17] Van Alsten, J.G., Coburn, J.C. (1994). Structural Effects on the Transport of Water in Polyimides. *Macromolecules*, 27, 3746–3752.

[18] Mensitieri, G., Lavorgna, M., Larobina, D., Scherillo, G., Ragosta, G., Musto, P. (2008). Molecular Mechanism of H_2O Diffusion into Polyimides: A Model Based on Dual Mobility with Instantaneous Local Nonlinear Equilibrium. *Macromolecules*, 41, 4850-4855.

[19] Ravji, S.H. (2015). *Mechanisms of water vapour transport in polyimide thin films for applications in humidity sensing*. Ph.D. Thesis, University of Glasgow, UK.

[20] Zent, A.P., *et al.* (2010). Initial results from the thermal electrical conductivity probe (TECP) on Phoenix, *J. Geophys. Res. Planets*, 115, E00E14, http://dx.doi.org/10.1029/2009JE003420.

[21] Tetelin, A., Pellet, C. (2006). Modeling and Optimization of a Fast Response Capacitive Humidity Sensor. *IEEE Sensors J.*, 6(3), 714–720.

[22] Wildmann, N., Kaufmann, F., Bange, J. (2014). An inverse-modelling approach for frequency response correction of capacitive humidity sensorsin ABL research with small remotely piloted aircraft (RPA). *Atmos. Meas. Tech.*, 7, 3059–3069.

[23] Rückerl, A., Huppmann, S., Zeisel, R., Katz, S. (2014). Monolithic integrable capacitive humidity sensing method for material characterization of dielectric thin films. *Microelectronics Reliability*, 54, 1741–1744.

[24] Sadaoka, Y. (2009). Capacitive-Type Relative Humidity Sensors with Hydrophobic Polymer Films. Ch.3, 109–152, in: Comini, E., Faglia, G., Sberveglieri, G. (eds.). *Solid state Gas Sensing.* New York: Springer Science & Business Media.

[25] Fratoddi, I., Bearzotti, A., Venditti, I., Cametti, C., Russo, M.V. (2016). Role of nanostructured polymers on the improvement of electrical respons-based relative humidity sensors. *Sens. Actuators B*, 225, 96–108.

[26] Żukowski, P., Kołtunowicz, T.N., Kierczyński, K., Subocz, J., Szrot, M. (2016). Formation of water nanodrops in cellulose impregnated with insulating oil. *Cellulose*, 22(1), 861–866.

[27] Cirmirakis, D., Demosthenous, A., Saeidi, N. (2013). Humidity-to-Frequency Sensor in CMOS Technology With Wireless Readout. *IEEE Sensors J.*, 13(3), 900–908.

[28] Wang, B., Law, M.K., Bermak, A. (2012). A Low-Cost Capacitive Relative Humidity Sensor for Food Moisture Monitoring Application. *Proc. of the 4th Asia Symposium on Quality Electronic Design.(ASQED)*, 509(2), 95–99.

[29] Datasheet SHT21. Humidity and Temperature Sensor IC. *https://www.sensirion.com* (May 2016).

www.czasopisma.pan.pl PAN www.journals.pan.pl

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 617–630.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

ANALYSIS OF PROPERTIES OF AN ACTIVE LINEAR GESTURE SENSOR

Krzysztof Czuszyński, Jacek Rumiński, Jerzy Wtorek

Gdańsk University of Technology, Faculty of Electronics Telecommunications and Informatics, G. Narutowicza 11/12, 80-233 Gdańsk, Poland (✉ krzysztof.czuszynski@pg.edu.pl, +48 58 347 1725, jacek.ruminski@pg.edu.pl, jerzy.wtorek@pg.edu.pl)

Abstract

Basic gesture sensors can play a significant role as input units in mobile smart devices. However, they have to handle a wide variety of gestures while preserving the advantages of basic sensors. In this paper a user-determined approach to the design of a sparse optical gesture sensor is proposed. The statistical research on a study group of individuals includes the measurement of user-related parameters like the speed of a performed swipe (dynamic gesture) and the morphology of fingers. The obtained results, as well as other *a priori* requirements for an optical gesture sensor were further used in the design process. Several properties were examined using simulations or experimental verification. It was shown that the designed optical gesture sensor provides accurate localization of fingers, and recognizes a set of static and dynamic hand gestures using a relatively low level of power consumption.

Keywords: optical sensor, gestures, mobile devices, Monte Carlo Simulation, smart glasses.

1. Introduction

The development of mobile devices and a wide range of their possible applications have stimulated extensive research on possible interaction methods, especially those related to the design of low-power gesture sensors. Touchless interfaces are of special interest as they could prove useful in some specific areas, *e.g.* in healthcare [1, 2]. Let us focus on basic optical sensors; to date they have been mainly used as supplementary input devices in mobile equipment like smartphones and tablets. Three types of power-effective basic optical gesture sensors can be distinguished regarding the number of light-sensitive elements (detectors) of the device.

The first type are one-detector sensors. They can detect a proximity level [3], a swipe event along a single axis and a dynamic pose of hand [4, 5]. To determine the swipe direction while only having one LED, the sensitivity gradient [5], light inhibitor [6] or asymmetric optical blocks [7] can be applied. Solutions involving two light sources are also used [8]. The second type are two-detector sensors. One- and two-detector sensors, which mainly recognize swipe events are called *Motion Gesture Sensors* (MGSs) [9, 10]. They can estimate the direction and often the speed of detected movement, but not a precise range of swipe. The third type are sensors with several detectors, excluding matrices of detectors, like RGB cameras. Kong *et al.* have proposed three-detector MGSs with [11] and without [12] optical blocks enclosed in a single chip, handling swipes in 3D. Withana *et al.* have proposed a modular sensor composed of photodiode-LED pairs [13], which can be arranged in a linear or triangular form. It can handle a wide variety of gestures but it is not able to localize a hand precisely. They have also described a two-emitter, six-receiver prototype sensor for virtual reality glasses, which detects swipes and push/pull gestures [14]. Chuang *et al.* have mounted emitter-receiver pairs in the corners of a mobile device that enabled to recognize a hand position basing on trilateration [15]. Tang *et al.* have proposed a linear, 10-detector IR transceiver, as an implementation of the

Article history: received on Feb. 08, 2017; accepted on May 21, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0052.

virtual computer mouse, which tracks a hand position and can detect three click gestures [16]. A similar solution was applied to multi-touch interactions with a mobile device, mounted at the edges of its housing [17]. The gestures based on tracking eyeball movements, recorded by 4 photodiode-LED pairs per eye, were examined as well [18].

Considering the gestures' taxonomy, MGSs mostly detect motion-related events (dynamic gestures) but also the presence of a still hand (static gesture). In *human system interactions* (HSI), both of these gesture types are discrete. This means that the system responds after a whole gesture is completed (or its duration exceeds a threshold). Some interfaces handle continuous gestures, *e.g.* [15]. In terms of HSI, the system navigated by such an interface reacts while the gesture is performed.

As indicated, basic optical gesture sensors often handle a narrow set of gestures, mainly using a discrete dynamic. But when more gestures can be directly associated with varied actions the system response becomes more rapid and more precise. Moreover, there are not many basic optical sensors, which can handle both continuous and wide sets of discrete gestures. Therefore, the intension of this work is to design a touchless, power-efficient, several-detector optical sensor, capable of handling a wide set of both discrete and continuous gestures. The aim of the paper is to adjust the parameters of the proposed construction of a sensor, so that it would be capable of detecting certain hand movements and differentiating various hand arrangements, while respecting user-defined spatiotemporal requirements. The study is based on simulations and experiments carried out with the use of actual implementation of the sensor.

The paper is organized as follows: the first section contains the introduction, an outline of the state-of-the-art technology and the objective of the work. The second section presents the methods applied to describe and measure the performance of the proposed optical sensor. The obtained results of simulations and experiments are included in the third section. The summary and conclusions are given in the last section.

2. Materials and methods

2.1. Proposed sparse optical gesture sensor

Basing on initial requirements and the present state-of-the-art, a prototype of the sparse construction of optical gesture sensor with 8 aligned IR *photodiodes* (PD) has been designed (Fig. 1a). The photodiodes are evenly distributed on an 8 cm long *printed circuit board* (PCB), with 4 IR LEDs. All optical elements are mounted on the same plane and face in one direction. The applied photodiodes are chips with built-in operational amplifiers TSL260RD and the used LEDs are KA-3528SF4S. The sensor's microprocessor, a 5 V supplied PIC24FV16KA302, is employed in data sampling, processing and duplex communication via a UART serial interface.

The distributions of PDs and LEDs on the PCB are denoted by d_{PD} and d_{LED}, respectively. In the design the LEDs are placed along an axis designated by photodiodes and each is separated from the closest PD by $d_{PD}/2$. A sensor is considered sparse when $d_{PD} \geq 5a$, where a is the side of the active area of a square-shaped PD. The angular parameters of optoelectronic elements are described by β_{PD} and β_{LED} – the field of view of elements (FOV) – which can be adjusted by the application of optical blocks of appropriate height (h_{bPD}, h_{bLED}). They can be modified together using a set of overlays created with the 3D printing technology. The angular sensitivity of PDs and LEDs is assumed to be a cosine function. The position of each element within the sensor is described in the Cartesian coordinate system, with the origin located between two middle PDs, (Fig. 1b). The position of the centre of active area of an optoelectronic element is defined by the coordinates of that element.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 617–630.

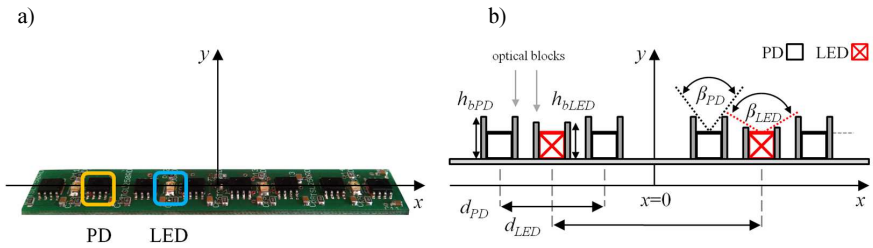

Fig. 1. An optical gesture sensor (without optical blocks for better clarity) (a). A sketch of the sensor descriptive parameters – a side view of the central part of the sensor (b).

A sparse optical gesture sensor analyses the pattern of light intensity obtained from the PDs. It can operate in two modes, depending on the *ambient light* (AL) level. In strong AL conditions, the pattern of shadow caused by a hand covering the light is analysed. In weak AL conditions, the pulses of sensor LEDs highlight a hand performing a gesture and the pattern of reflected light is analysed. These modes are called the passive and active ones, respectively. In this paper only the active operating mode is considered. The pattern is to be produced by the shadow created by the hand performing a gesture; it indicates the arrangement of fingers. Two types of pose are considered. *Spread* (S) indicates that all of the *fingers* (F) involved in the gesture are separated, in opposition to the *Joined* (J) fingers' arrangement. The codes of gestures are presented in Table 1.

Table 1. The codes of static discrete gestures to be handled by the sparse optical sensor.

Fingers involved	**1**	**2**	**3**	**4**
Spread arrangement	1FS	2FS	3FS	4FS
Joined arrangement	–	2FJ	–	4FJ

2.2. User studies

The research on a control group of individuals was performed in order to gain referential requirements for the sensor design in terms of physical dimensions and throughput of the system. The group consisted of 41 Caucasian volunteers (21 females, 20 males, age: 26.4 ± 6.1 years).

In the first part of the experiment the following parameters were measured: R, a radius of an index finger (1FS), $D2$, a width of an index and middle fingers joined (2FJ) and L, a spacing between the centres of middle and index fingers while freely arranged (2FS).

In the next part, the volunteers were asked to perform three series of swipe gestures along the sensor, at a distance of about 1−5 cm. Each series: 2FJ slow swipe, 2FJ fast swipe and 2FS fast swipe consisted of 5 repetitions. The gestures were recorded with a referential sampling frequency, $f_r = 2$ kHz, in order to precisely measure the velocity of the movements. The signals were sampled directly from the output of photodiode chips using a USB-1608GX DAQ device. The optical sensor was set to full light mode ($D = 100\%$).

2.3. Sensor design features

The sensor's abilities of recognition and handling of different gestures result from its properties. They have been described by a set of parameters, which are discussed and evaluated in the following subsections.

2.3.1. Illumination pattern flatness

Let the numbers of photodiodes and LEDs, n_{PD}, n_{LED} respectively, and both d_{PD}, d_{LED}, define a specific formation of optical elements within a sparse gesture sensor. Three different formations were initially examined in this paper (Fig. 2).

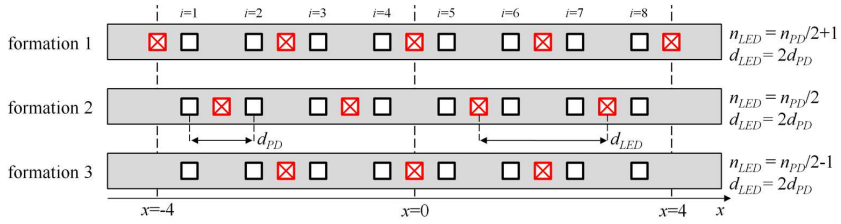

Fig. 2. Three types of formations of optoelectronic elements (top view) within a sparse linear sensor. Optical blocks are not presented for the sake of clarity of the figure.

Each of them creates its own illumination pattern differing in flatness along x and y axes. The greater the $flatness_y$ (on the y axis), the greater the effective range of the sensor (photodiodes saturate further from the device). A greater $flatness_x$ (along x axis) enables a more linear estimation of the hand position when no correction is applied. Consider a flat, w cm wide, obstacle positioned centrally above the most inner and most outer photodiodes at a distance h from the sensor in two separate measurements. The $flatness_x$ is calculated as a ratio of the strengths of signals from the most outer (second measurement) and most inner (first measurement) photodiodes (in the case of n_{PD} = 8, 1st/8th and 4th/5th respectively). The $flatness_y$ is expressed as a $flatness_x(h)$ function.

2.3.2. Operating area

Kim *et al.* have described their two-detector sensor with three zones [9]. An obstacle (hand/fingers performing a gesture) is in the dead zone when any of the detectors can see it. In a linear sensor, considering a 2D section, there are n_{PD} + 1 triangular dead zones (Fig. 3). However, in the active operating mode, the obstacle has to be within the FOV of both detectors and light sources. Therefore, the illumination system has its own dead zone as well. The heights of the dead zones of detectors and the illumination system can be respectively described by:

$$h_{ddz} = d_{PD} / \left(2tg(\beta_{PD}/2)\right), \qquad h_{idz} = d_{LED} / \left(2tg(\beta_{LED}/2)\right). \tag{1}$$

When an object is located closer than at the operating distance $h_{op} = \min(h_{ddz}, h_{idz})$, its visibility depends on its location along the x axis. In this area indistinct results are obtained, hence it is called an ambiguous zone and it has to be minimized. The obstacle delivers proper data for gesture recognition if located in the detectable zone, at $h \geq h_{op}$. A width of the detectable zone, l (Fig. 3), at height h, for a sparse linear sensor composed of n_{PD} photodiodes can be described as:

$$l(h) = (n_{PD} - 1)d_{PD} + 2h \cdot tg(\beta_{PD}/2). \tag{2}$$

The sensor is dedicated to detection of close range gestures. Hence, the upper range limit is assumed to be h_{max} = 5 cm, since interference with objects located further away is undesirable. As the sensor is touchless, the minimum operating distance considered, where contact with the device could be avoided, is h_{min} = 1 cm, hence $h_{op} \leq h_{min}$.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 617–630.

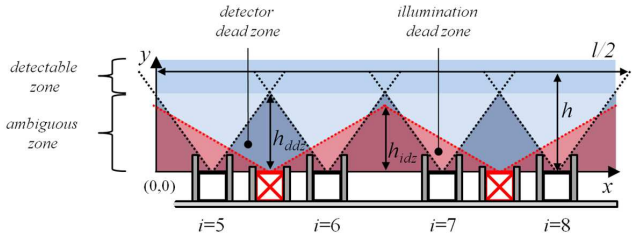

Fig. 3. A sketch of the right side of the sparse optical sensor with three types of zones marked.

2.3.3. Resolving power

An important property of the sensor is its ability to differentiate FS and FJ arrangements (*e.g.* 2FS vs 2FJ). Therefore, a distance at which the sensor detects an indentation in the pattern of reflected light intensity produced by a gap in 2FS, should be maximized. In the 2FS, fingers of radius R are separated from each other by L (Fig. 4a). The depth of indentation, I_D, in an observable pattern (Fig. 4b) is the separation criterion. Lack of indentation means that L of fingers in 2FS at h is too small for a given set of parameters of the sensor. Therefore, the impact of d_{PD} and β_{PD} on the resolving power of the sensor along two axes is considered. The *shift* parameter is a distance between the centre of symmetry of a sensor ($x = 0$) and the x component of the centre of symmetry of a given fingers' arrangement system (Fig. 4a). The variability of I_D as a function of *shift*, h for fixed R and L values is of interest. The I_D is obtained from:

$$I_D = 1 - \frac{v_M}{\min(v_L, v_R)}, \tag{3}$$

where v_L and v_R are values of peaks adjacent to the common middle value v_M from the left and right sides, respectively (Fig 4b).

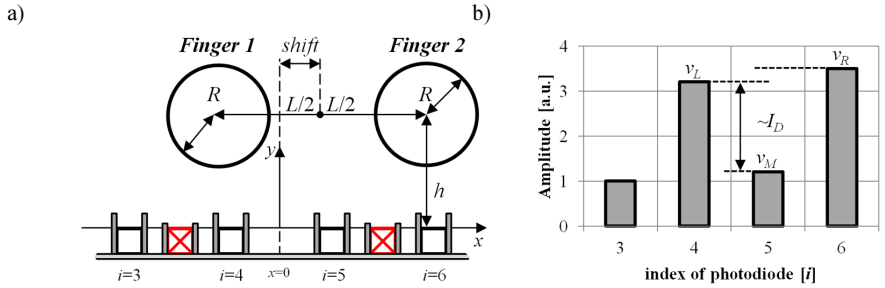

Fig. 4. A cross-section of the middle part of the sensor and fingers in the 2FS arrangement (a). A corresponding light intensity pattern with the meaningful values for calculation of indentation depth, I_D (b).

2.3.4. Spatial and temporal sensitivity

Motion detection describes the minimal distance covered by a hand which causes a detectable swipe at a distance h [11]. In the sparse sensor, the spatial sensitivity can be defined as the minimal noticeable displacement of an obstacle moving along the x axis. The movement is considered as slow enough so that the velocity is not an influencing factor. A current position of the centre of gravity of a symmetric (along the x axis) obstacle located in front of the sensor, x_o, can be estimated with the formula:

$$x_o = \sum_{i=1}^{n_{PD}} (v_i \cdot x_i) / \sum_{i=1}^{n_{PD}} v_i \,, \tag{4}$$

where v_i is a light intensity value sampled by an i-th photodiode and x_i is the PD's position in the Cartesian coordinate system. The function of the real position of an obstacle along the x axis vs x_o is of interest. Different obstacles and distances, h, should be examined.

The motion detection can be also considered in the time domain. The sampling frequency, f_s, of the sensor defines the upper limit of velocity of the fastest noticeable swipe, V_{not}. It is termed a temporal sensitivity. A sparse linear sensor has to perform at least double sampling during a hand swipe to estimate the velocity of the movement. In the most favourable boundary case, the sensor operating at f_s, performs the sampling at the moment when a hand moving with V_{not} appears above the first and the last photodiodes in the following sampling cycles. The double of f_s ensures the double sampling also in the least favourable case. Therefore, the minimal f_s of the sensor may be expressed as follows:

$$f_s = 2V_{not} / \left((n_{PD} - 1) \cdot d_{PD} \right). \tag{5}$$

As f_s affects power consumption levels it has to be based upon moves performed by the user.

2.3.5. Power consumption

Such parameters as brightness of the LEDs, n_{LED}, a target f_s of the device and a fill factor, D, determine the power consumption of the illumination system of the sensor. D is a duration of the LEDs' turned-on state within the duration of the sampling period. Its value depends on the settling time of applied photodiode chips and the ADC's sampling period duration.

2.4. Monte Carlo simulations

In this work simulations were used to determine different geometrical configurations, sensor element formations and individual parameter adjustment of a virtual optical gesture sensor instead of building many versions of the physical device. Considering the light-solid interactions, employing the *Monte Carlo method-based simulations* (MCMS) is a commonly applied approach [19]. In our study, a simplified interaction model, described in [20], was used. At the initialization stage each photon is given a weight, W, which decreases according to the length of the ray and reflection events. The unitless weight is an equivalent of the amount of energy of a photon. A fixed number of photons, N = 50 million, take part in each simulation. They are generated by the light sources of the modelled sensor, interact with an obstacle and possibly hit one of the detectors. An obstacle can be described by 4 parameters: x_o, y_o, R, and w. A point (x_o, y_o) indicates the position of the centre of obstacle in the described Cartesian coordinate system. R is a radius of the curved part of a round obstacle (*e.g.* finger), whereas the w parameter describes a width of the plane part of the modelled obstacle.

The model solver was designed in Matlab. However, a large number of sampling events led us to implement it within a multithread C# application, which has accelerated the computation approximately 75 times.

2.5. Laboratory experiments

The laboratory experiments were designed and performed to determine the sensor properties and to verify the light interaction model. For this purpose, the sensor described in Subsection 2.1, configured as d_{PD} = 1 cm, β_{PD} = 60° and d_{LED} = 2 cm, β_{LED} = 120°, has been used. During the experiments, the examined light reflecting obstacle was attached to a trolley, which was

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 617–630.

moving along a straight track at a velocity of 5 cm/s. An *analogue-to-digital converter* unit (ADC) of the microprocessor sampled the signals from photodiodes into 12-bit, 4-digit numbers at a rate of 40 Hz. The output was read by the PC via a UART interface.

2.6. Correspondence between simulations and measurements

The proposed light-solid interaction model was verified. In the first stage the amplitude vs distance relationship was checked. In the experiment, a flat white cardboard was set to move away from the sensor, so a point P travelled a perpendicular path from h_1 to h_2 (Fig. 5a). The trial was reproduced in the simulation environment. The unit-less values from virtual photodiodes were adjusted to match the measured ones, expressed in volts, by minimizing the residual sum of squares (RSS). The objective parameter was a multiplication factor. The amplitudes were averaged from two middle photodiodes of the sensor ($i = 4$, $i = 5$).

In the next step, the angular sensitivity of the virtual sensor was verified. Two different obstacles were set to move along the sensor, at a given h, so a point P travelled a parallel path from x_L to x_R (Fig. 5a). The passing of an obstacle produced light reflection patterns recorded by each PD. Let S^i and M^i be k-sample long vectors with reflection patterns (windowed from the recorded PD signal) from an i-th photodiode, obtained from simulations and measurements, respectively. It is assumed that the pattern associated with the passing is not wider than k samples. S^i and M^i vectors are normalized over max(S) and max(M), respectively. They are compared using the proposed distance parameter:

$$\zeta^i = \sum_{j=s}^{f} \left| (S_j^i - M_j^i) / \max(M_j^i) \right| \cdot k / (f - s), \tag{6}$$

where: k is a length of the analysis window; s and f are indices, which create an analysis sub-window. The sub-window starts/finishes at an index j where the greater of S^i, M^i vectors' values reaches/falls behind the v_{min} (2% of max(M^i)) (Fig. 5b). The value of v_{min} was chosen experimentally. The application of the sub-window makes the comparison of results of narrow and wide shapes relevant when using (6).

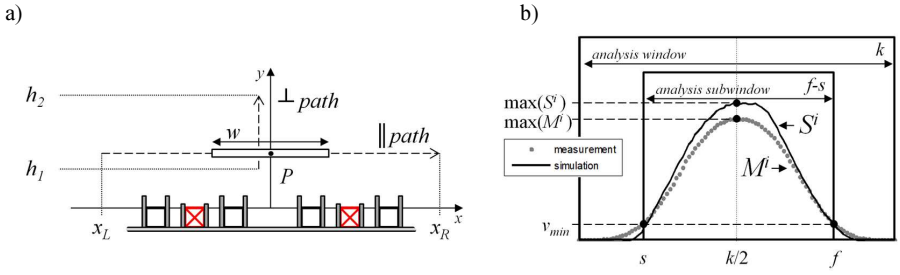

Fig. 5. Parallel and perpendicular paths of an obstacle in relation to the sensor (a). A view of windowed reflection patterns for comparative analysis (idea of the subwindow) (b).

3. Results

3.1. Results of user studies

Distributions of parameters obtained from the control group are presented in Table 2 and in Fig. 6. The results obtained from the control group were applied as the referential data in the research on the related sensor's parameters.

Table 2. Statistical parameters of the control group.

PARAMETER	MEDIAN FEMALES	MEDIAN MALES	MEDIAN TOTAL	STDEV FEMALES	STDEV MALES	STDEV TOTAL
R [cm]	0.7	0.85	0.75	0.05	0.10	0.11
D2 [cm]	2.9	3.6	3.2	0.18	0.30	0.41
L [cm]	3.1	4.0	3.6	0.52	0.65	0.76
AGE [yrs]	23	25.5	24	3.72	7.64	6.11

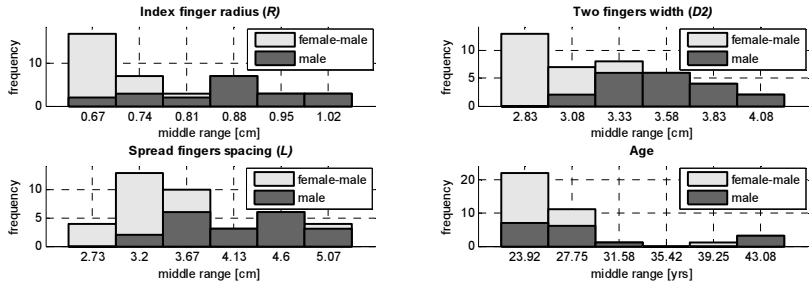

Fig. 6. Characteristics of the control group in a form of histograms.

Perception of speed is a subjective matter. Therefore, the most different values were rejected in order to present compact histograms (Fig. 7) but were included in the statistics (Table 3). Velocities of fast swipes of 2FJ and 2FS were taken together as 2Fx.

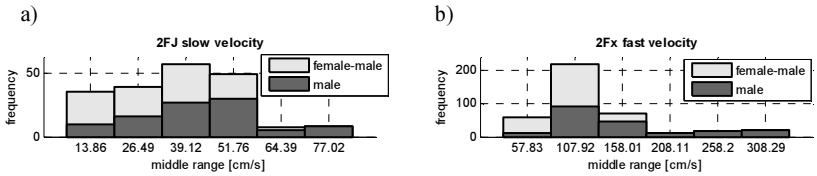

Fig. 7. Distribution of swipe gestures' velocities: 2FJ slow (a); 2FJ and 2FS fast (b).

Table 3. Statistical parameters of gestures performed by the control group.

Gesture	MEDIAN FEMALES	MEDIAN MALES	MEDIAN TOTAL	STDEV FEMALES	STDEV MALES	STDEV TOTAL
2FJ slow [cm/s]	33.90	43.02	37.04	15.85	20.49	19.08
2Fx fast [cm/s]	95.24	125	105.26	37.42	110.53	87.47

3.2. Correspondence between simulations and measurements

The simulations were validated in an amplitude vs distance (h) test using a flat white cardboard of w = 5 cm, to ensure uniform illumination of the middle PDs. An h value varied from 3 to 10 cm. The obtained RSS was equal to 0.48. The formula for the trend function (dashed plot) confirms compatibility of the results with the inverse power law model (Fig. 8).

In the experimental part of angular sensitivity correspondence test, the trolley with an attached obstacle moved 20 times from $x_L = -20$ cm to $x_R = 20$ cm. For each PD a model light intensity pattern, produced by a single passing of an obstacle along the sensor, was arbitrarily chosen. Then, 10 other patterns, most similar to the model pattern upon the correlation coefficient, were selected and averaged, creating M^i vectors. The formula (6) was applied for

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 617–630.

two white obstacles: a cardboard and a cylinder, as described in Table 4. The table contains the results of how the model corresponds with the reality.

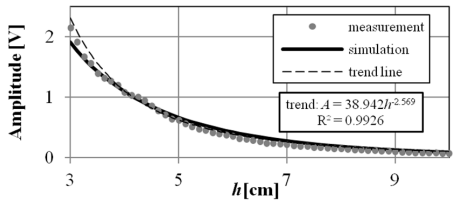

Fig. 8. The amplitude vs distance (h) relationship – real and obtained from the simulated sensor.

Table 4. The distance parameter describing a degree of similarity of simulations and measurements.

photodiode [*i*]	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
ζ^i (cardboard, w = 5 cm, h = 4 cm)	7.09	5.70	3.11	1.79	1.86	4.11	2.01	4.43
ζ^i (cylinder, R = 0.75 cm, h = 2.8 cm)	9.76	9.20	5.68	4.87	5.01	5.28	7.20	7.28

Figure 9 presents the best fitting pair (i = 4) for the cardboard obstacle case (a) and the worst fitting pair (i = 1) for the cylinder obstacle case (b).

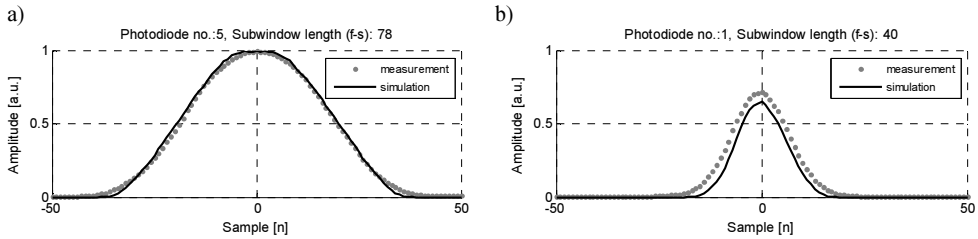

Fig. 9. Comparison of shapes obtained by simulations and measurements in k-sample windows.

3.3. Optical sensor characteristics

The correspondence level of the results obtained from the real world implementation and virtual sensors is high. It enables to accomplish research based on the optical sensor features, which is difficult to perform experimentally, in the elaborated simulation environment.

3.3.1. Illumination pattern

N photons are emitted in simulation by the illumination system of each of the considered formations of the sensor (Fig. 2). Their illumination efficiency can be compared, since the same amount of power is dissipated. The considered flat obstacle was imitating a 2FJ arrangement, w=3.2 cm (median $D2$ from Table 2). In order to narrow the illumination pattern flatness problem, the following parameters were fixed: β_{PD} = 60°, β_{LED} = 120°, n_{PD} = 8 and d_{PD} = 1 cm. Strengths of signals from individual PDs at h = 1 cm produce an illumination pattern (Fig. 10a). The inter-formation highlight efficiency was examined for middle photodiodes of the sensor (Fig. 10b). A single value of $flatness_x$ refers to a given formation of elements and its distance from the sensor, while $flatness_x(h)$ enables to observe flatness of the illumination pattern along both x and y axes (Fig 10 c).

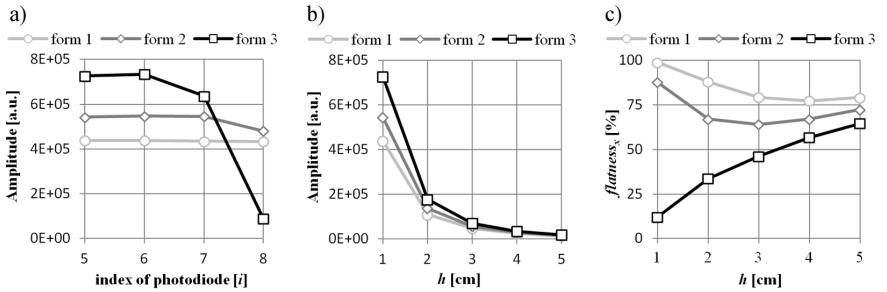

Fig. 10. Strength of the reflected light signals with a 2FJ positioned over a given photodiode at h = 1cm (a). Strength of the reflected light signals from i=5 photodiode as a function of h (b). $Flatness_y$ functions composed of individual $flatness_x$ values (c).

3.3.2. Operating area

A height of the ambiguous zone, h_{op} , has to be kept below h_{min}. With $d_{LED} = 2d_{PD}$, $\beta_{LED} = 120°$ and β_{PD} up to 82° (1), h_{op} is defined by geometry of the detector. The relation d_{PD} vs β_{PD}, based on (1), helps to fulfill the condition $h_{op} \leq h_{min}$ (Fig. 11a, painted area). $l(h)$ depends on a triple: n_{PD}, d_{PD} (cm) and β_{PD}. The value of $l(h)$ was calculated for differently configured sensors, described by 3 triples: (6, 1.4, 80°), (8, 1, 60°) and (10, 0.8, 45°), respectively (Fig. 11b). The values of n_{PD} and d_{PD} from each triple were selected so their products, which mean the separation between the boundary PDs, $l(0)$, are possibly equal. d_{PD} and β_{PD} from the triples respect the relation: $h_{op} \leq h_{min}$. A smaller d_{PD} is not considered due to the size of the photodiode chip.

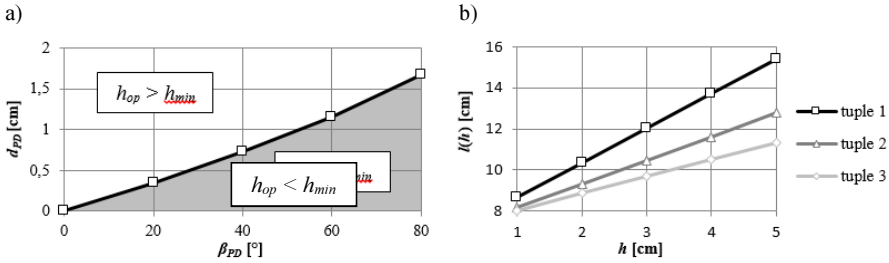

Fig. 11. A function binding values of β_{PD} and d_{PD} to keep h_{op} at a proper level (painted area) (a). The values of $l(h)$ for three different configurations of the sensor, described by three triples (b).

3.3.3. Resolving power

Dependencies of the indentation depth, I_D, were obtained from the simulations. The values of parameters of a 2FS gesture, R and L, corresponded with median values from Table 2. Three virtual sensors described by 3 triples (Subsection 3.2.2), with β_{LED} = 120°, each with a relative placement of PDs and LEDs as in formation 1, were examined. Notice that the 1st and 3rd triple sensors have no LED at the origin in this case (Fig. 1). The $I_D(shift)$ functions at half of the range (h = 3 cm) were calculated upon the obtained simulation results (Fig. 12a). The smallest of local minimums, which are measured between adjacent bulges of the I_D function (marked by a circle in Fig. 12a), is considered as the minimal I_D at a given h, when plotting the min($I_D(shift)$) vs h function (Fig 12b).

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 617–630.

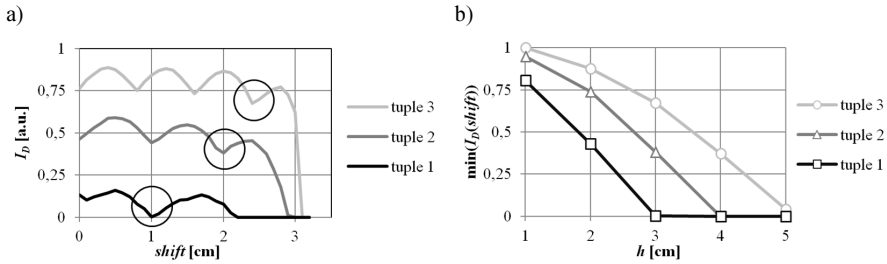

Fig. 12. The $I_D(shift)$ functions calculated for a 2FS located at h = 3 cm for three triples of parameters (a). min($I_D(shift)$) vs h functions for three triples of parameters (b).

3.3.4. Spatial and temporal sensitivity

Three obstacles were considered in the simulation research on spatial sensitivity. The 1FS one was a round obstacle of R = 1.7 cm, whereas 2FJ and 4FJ ones were flat planes of w equal to 3.2 and 6.4 cm (doubled *D2*) respectively. In each trial, the centre of gravity of an obstacle, x_o, was located in a virtual space at (0, h) and the movement along the x axis was simulated, so the *shift* has changed. The virtual sampling was performed after every 0.1 cm of actual *shift*. The calculated *shift* of each obstacle, perceived by the virtual sensor configured as in the 3rd triple (Subsection 3.2.2) of formation 1, at h = 3 cm was obtained upon (4) (Fig. 13a). The standard deviation of the position for different h values was examined for the 2FJ obstacle. It was calculated for *shift* values in ranges of 0–0.5 cm, 0–1 cm up to 0–4 cm (Fig 13b). The 2FJ arrangement was selected as it reflects more light than 1FS and is more precise than 4FJ.

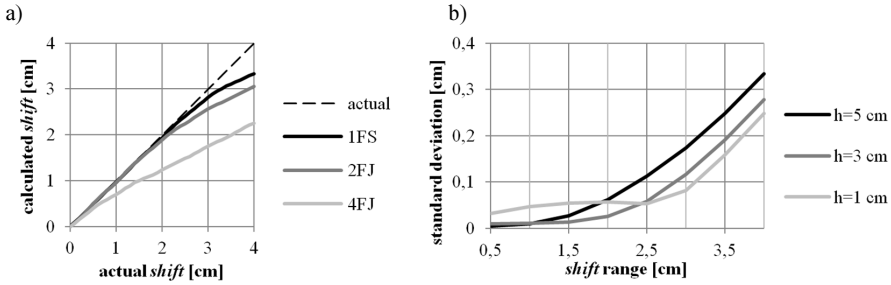

Fig. 13. Calculated *shift* vs actual *shift* for different arrangements at h = 3 cm (a). Standard deviation of position along x axis for 2FJ as a function of h (b).

Table 5 presents the relationship between f_s and V_{not} for the sensor configured according to triple 2 (as the real world implementation). Due to the possibly equal inter-triple products of n_{PD} and d_{PD} the numbers in Table 5 would be at least very similar for other configurations of the sensor. In the summary, the total of 4 cases were examined. Case 1 shows V_{not} for the highest possible f_s value of the device, respecting the required sampling and signal settling times. In cases 2 and 3 V_{not} was selected as the maximal and median velocity of fast swipes, respectively (2Fx fast swipe from the control group). Case 4 shows the noticeable swipe speed for a selected sampling frequency of the gesture sensor, f_s = 40 Hz. Notice how f_s affects the D value and hence the power consumption. The C parameter indicates the percentage of noticeable fast swipes among all of the fast swipes performed by the control group, when a given f_s is applied. Hence, the selected frequency f_s is a compromise between C and D.

Table 5. Relations between a sampling frequency and a detectable swipe velocity.

CASE	GIVEN *Fs* [Hz]	NOTICEABLE SPEED [cm/s]	GIVEN SPEED [cm/s]	REQUIRED *Fs* [Hz]	*C* [%]	*D* [%]
1	2666.67	9333.33	–	–	100	100
2	–	–	666.67	185.19	100	7.14
3	–	–	105.26	29.24	53.4	1.13
4	40	140	–	–	74.9	1.5

3.3.5. Power consumption

Using the real world implementation of the sensor (Fig. 1a, triple 2) the total time required for sampling of 8 channels (LEDs' turned-on period) was measured to be 375µs. Therefore, complying with the selected sampling frequency, f_s = 40 Hz, the resulting D value is 1.5%. In the illumination system of the used sensor 4 LEDs were applied, each consuming a peak current I_p = 30.83 mA. Hence, an illumination system powered by a 123.32 mA peak current consumes the total of 1.85 mA.

4. Discussion and conclusions

Some initial, preliminary results of the examined construction of a sparse sensor were already presented in [7, 21]. However, a more complex analysis of its parameters is presented in the paper.

The simulations reveal significant variation of illumination uniformity at h = 1 cm within formations (Fig. 10a). The inter-formation variability of illumination strength, in an example of one of central photodiodes of the sensor, drops with an increase of h value (Fig. 10b). The analysis indicates formation 1 as the one producing the most flat illumination pattern along both axes (Fig. 10c). It is characterized by very high $flatness_x$, especially in an area close to the sensor (h < 3 cm). Such a sensor also saturates with a reflecting obstacle located closer to the device in comparison with other formations (a wider effective range).

The reported operating distal range of optical sensors often reaches tens of centimetres [4, 11, 14]. However, the proposed sparse sensor is considered to handle unobtrusive gesticulation, performed at a close distance, hence the upper range h_{max} = 5 cm, similar to [17], was assumed to be sufficient. As presented (Fig. 8), the strength of the signal at h_{max} for a 5 cm wide obstacle is still significant, but it saturates at a distance $h > h_{min}$. The strength of the signal depends on the width of the reflecting object, so that the power of the LEDs needs to be adjusted.

The operating area can be controlled by the height of an ambiguous zone ($d_{PD} - \beta_{PD}$ relation) but the parameters of the sensor have to be selected carefully (Fig. 11a).

The resolving power of the sparse sensor enables to differentiate static hand gestures. This is an advantage in comparison with the solutions where the fingers' arrangement can be recognized based only on dynamic gestures [9, 10]. The I_D functions indicate that the sensor configured as the 2nd triple (8, 1, 60°) has a sufficient resolving power at h = 3.5 cm, if a threshold I_{Dt} = 10% is applied. However, the 3rd triple (10, 0.8, 45°) sensor can see FS arrangements almost in the whole considered range ($h \leq h_{max}$). The selectivity of individual PDs (β_{PD}) and the sparse construction of the sensor enable rough differentiation between the widths of light reflecting obstacles; hence also the distinguishing 1FS, 2FJ and 4FJ gestures is plausible.

Common MGSs need a clear swipe in order to detect a hand movement, *e.g.* a 2 cm long one [10]. However, typical interfaces, which are oriented towards the continuous gestures, are able

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 617–630.

not only to notice a swipe but also to estimate a current location of a hand/finger. The position standard deviation errors in one such interface are equal to 0.03 cm, 0.02 cm, 0.01 cm in the x, y and z axes, respectively [15]. The accuracy of hand localization in the proposed sparse sensor was examined only along the x axis. The results show that for the 2FJ fingers' arrangement the standard deviation of the calculated position is the smallest when the hand operates above the central part of the sensor. In practice, it barely exceeds 0.1 cm for fingers moving within the middle 5 cm of the sensor range ($x_L = -2.5$ cm $x_R = 2.5$ cm), regardless of h (when h value is within the operating distance, $h_{min} \leq h \leq h_{max}$) (Fig. 13b). Therefore, the 2FJ should be considered as the main arrangement for continuous gestures. The obtained resolution, for a basic sensor, which is oriented on both continuous and discrete gestures, is promising but correction methods and localization on the z axis have to be examined.

Considering the detection of swipes of equal velocity, the sensor with a sparse construction can operate roughly at a $(n_{PD}-1)d_{PD}$ times lower f_s in comparison with a single-chip device. Therefore, due to a relatively low sampling rate the current consumption is lower than that reported for many optical sensors in the literature, which can be found in a range of 10-20 mA [4, 9, 10, 11] but 3.78 mA sensors were reported as well [12] (all solutions with 1 LED, enclosed in a single chip). The power consumption was found to be around 8 mW [13, 14], but many sensors consume more than 20 mW [8, 13]. The proposed sparse linear sensor draws 1.85 mA (9.25 mW) for the illumination system. A single PD chip consumes around 20 μA when supplied only during the sampling period. Therefore, the total consumption related with the optoelectronic elements is 2.02 mA (10.1 mW) and 2.16 mA (10.19 mW) for systems with 8 PDs + 4 LEDs (a real world sensor) and 10 PDs + 6 LEDs (the best configuration of a virtual sensor), respectively. The total consumption can be further reduced by application of PDs with shorter settling times.

The proposed construction of sensor delivers two features in terms of gesture recognition: recognizing arrangements of individual fingers formed by a hand [7] and precise localizing a hand in relation to the sensor, regarding two axes. Those features enable to define sets of discrete and continuous gestures [22]. Therefore, basing on the results, the expected interactions between the user and the sensor are as follows: hand swipe events (along 1 axis), movements towards and from the sensor, mouse-like navigation (continuous) and combinations of static hand pose. A low power consumption of the proposed sparse gesture sensor makes it a promising solution for mobile devices. A rich set of gestures handled by the sensor enables to consider it as the main interface to adapted mobile operating systems. The size and shape of the device make it an attractive solution for smart glasses, since it can be easily mounted at a side of the frame as it is usually done [23, 24]. The advantages of the proposed gesture sensor enable to use it within special applications, *e.g.* in the industrial or sterile environment.

Acknowledgements

This work has been partly supported by NCBiR, FWF, SNSF, ANR and FNR in the framework of the ERA-NET CHIST-ERA II, project *eGLASSES – The interactive eyeglasses for mobile, perceptual computing,* and by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology.

References

[1] Czuszyński, K., Rumiński, J., Kocejko, T., Wtorek, J. (2015). Septic safe interactions with smart glasses in health care. *Proc. of EMBC 2015 Conference*, *IEEE Xplore*, 1604–1607.

[2] Mentis, H.M. (2015). Voice or Gesture in the Operating Room. *Proc. of CHI EA '15 Conference*, *ACM*, 773–780.

[3] Hinckley, K., Pierce, J., Sinclair, M., Horvitz, E. (2000). Sensing Techniques for Mobile Interaction. *Proc. of UIST '00 Symposium ACM*, 91–100.

[4] Metzger, C., Anderson, M., Starner, T. (2004). FreeDigiter : A Contact – free Device for Gesture Control. *Proc. of ISWC '04 Symposium*, *ACM*, 18–21.

[5] Manabe, H. (2013). Multi-touch gesture recognition by single photoreflector. *Proc. of UIST '13 Symposium, ACM*, 15–16.

[6] Gao, Y., Broga, A.M., Krishnaswamy, P. (2015). Contactless gesture recognition with sensor having asymmetric field of view. Patent no. EP 2866124 A1.

[7] Czuszynski, K., Ruminski, J., Wtorek, J., Vogl, A., Haller, M. (2015). Interactions using passive optical proximity detector. *Proc. of HSI 2015 Conference, IEEE Xplore*, 180–186.

[8] Cheng, H., Chen, A.M., Razdan, A., Buller, E. (2011). Contactless Gesture Recognition for Mobile Devices. *MIAA*.

[9] Kim, Y.S., Baek, K. (2013). A motion gesture sensor using photodiodes with limited field-of-view. *Optics Express*, 21(8), 555–560.

[10] Kong, K., Kim, Y.S., Kim, J.E., Kim, S., Baek, K. (2013). Single-Package Motion Gesture Sensor for Portable Applications. *IEEE Transactions on Consumer Electronics*, 59(4), 848–853.

[11] Kim, J.S., Yun, S.J., Seol, D.J., Park, H.J., Kim, Y.S. (2015). An IR Proximity-Based 3D Motion Gesture Sensor for Low-Power Portable Applications. *IEEE Sensors Journal*, 15(12), 7009–7016.

[12] Kim, J.S., Yun, S.J., Kim, Y.S. (2016). Low-power motion gesture sensor with a partially open cavity package. *Optics Express*, 24(10), 10537–10546.

[13] Withana, A., Peiris, R., Samarasekara, N., Nanayakkara, S. (2015). zSense : Enabling Shallow Depth Gesture Recognition for Greater Input Expressivity on Smart Wearables. *Proc. of CHI '15 Conference, ACM*, 3661–3670.

[14] Withana, A., Ransiri, S., Kaluarachchi, T., Singhabahu, C., Shi, Y., Elvitigala, S., Nanayakkara, S. (2016). waveSense : Ultra Low Power Gesture Sensing Based on Selective Volumetric Illumination. *Proc. of UIST '16 Symposium, ACM*, 139–140.

[15] Chuang, C., Chang, T., Jau, P., Chang, F. (2014). Applying the Kalman Filter to the Infrared-Based Touchless Positioning System with Dynamic Adjustment of Measurement Noise Features. *Proc. of IMPACT 2014 Conference, IEEE Xplore*, 84–87.

[16] Tang, S.K., Tseng, W.C., Luo, W.W., Chiu, K.C., Lin, S.T., Liu, Y.P. (2011). Virtual Mouse: A Low Cost Proximity-Based Gestural Pointing Device. *Proc. of HCI 2011Conference, Springer*, 491–499.

[17] Butler, A., Izadi, S. Hodges, S. (2008). SideSight: multi-'touch' interaction around small devices. (2008). *Proc. of UIST '08 Symposium, ACM*, 201–204.

[18] Lewandowski, T., Augustyniak, P. (2010). The System of a Touchfree Personal Computer Navigation by Using the Information on the Human Eye Movements. *Proc. of HSI 2010 Conf., IEEE Xplore*, 674–677.

[19] Zhu, C., Liu, Q. (2013). Review of Monte Carlo modeling of light transport in tissues Review of Monte Carlo modeling of light transport. *Journal of Biomedical Optics*, 18(5), 1–12.

[20] Czuszynski, K., Ruminski, J., Polinski, A. Bujnowski, A. (2016). Estimation of the amplitude of the signal for the active optical gesture sensor with sparse detectors. *Proc. of HSI 2016 Conference, IEEE Xplore*, 483–489.

[21] Bujnowski, A., Czuszynski, K., Ruminski, J., Wtorek, J., McCall, R., Popleteev, A., Louveton, N., Engel, T. (2015). Comparison of active proximity radars for the wearable devices. *Proc. of HSI 2015 Conference, IEEE Xplore*, 158–165.

[22] Czuszynski, K., Ruminski, J., Bujnowski, A., Wtorek, J. (2016). Semi complex navigation with an active optical gesture sensor. *Proc. of UbiComp '16 Conference, ACM*, 269–272.

[23] Amft, O., Wahl, F., Ishimaru, S., Kunze, K. (2015). Making Regular Eyeglasses Smart. *IEEE Pervasive Computing*, 14(3), 32–43.

[24] Serrano, M., Ens, B., Irani, P. (2014). Exploring the Use of Hand-To-Face Input for Interacting with Head-Worn Displays. *Proc. of CHI '14 Conference, ACM*, 3181–3190.

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 631–644.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

LEAK DETECTION IN WATERWORKS: COMPARISON BETWEEN STFT AND FFT WITH AN OVERCOMING OF LIMITATIONS

Aimé Lay-Ekuakille[1)], Giuseppe Griffo[1)], Paolo Visconti[1)], Patrizio Primiceri[1)], Ramiro Velazquez[2)]

1) University of Salento, Department of Innovation Engineering, Via Monteroni, 73100 Lecce, Italy (✉ aime.lay.ekuakille@unisalento.it, +39. 0832 297 821 822, giuseppe.griffo@unisalento.it, paolo.visconti@unisalento.it, patrizio.primiceri@unisalento.it)

2) Panamerican University, Faculty of Engineering, Av. Josemaría Escrivá de Balaguer 101, 20290 Aguascalientes, Mexico (rvelazquez@ags.up.mx)

Abstract

Detection of leakages in pipelines is a matter of continuous research because of the basic importance for a waterworks system is finding the point of the pipeline where a leak is located and – in some cases – a nature of the leak. There are specific difficulties in finding leaks by using spectral analysis techniques like FFT (*Fast Fourier Transform*), STFT (*Short Term Fourier Transform*), *etc*. These difficulties arise especially in complicated pipeline configurations, *e.g.* a zigzag one. This research focuses on the results of a new algorithm based on FFT and comparing them with a developed STFT technique. Even if other techniques are used, they are costly and difficult to be managed. Moreover, a constraint in the leak detection is the pipeline diameter because it influences accuracy of the adopted algorithm. FFT and STFT are not fully adequate for complex configurations dealt with in this paper, since they produce ill-posed problems with an increasing uncertainty. Therefore, an improved Tikhonov technique has been implemented to reinforce FFT and STFT for complex configurations of pipelines. Hence, the proposed algorithm overcomes the aforementioned difficulties due to applying a linear algebraic approach.

Keywords: uncertainty, measurements, magnetic sensors, leak detection in pipelines, regularization.

1. Introduction

Spectral analysis techniques have stimulated great interest in the fields of measurements and sensing systems. They are simple to apply and belong to software techniques based on computations as opposed to the hardware ones based on dedicated instrumentation. In this paper, we implement FFT and STFT algorithms, which is an obvious but not trivial approach when applied to a zigzag hydraulic circuit representing a complicated case-study since, in many cases, we do not refer to this kind of configuration but it is possible to find it in industrial applications. Leak detection in pipelines and waterworks has an economic impact on a budget of public and private company managing such services. Quality of water is associated with environment protection because of possible intrusion of pollution in pipelines and diminishing of water quality with pollution of aquifers. Leaks are considered as small pressure discontinuities in the original pressure trace and increase the damping of the overall pressure signal [1]. Such partial reflections divert energy away from the main waveform and increase the decay rate of the transient signal. The behaviour of this pressure trace is, therefore, an indication of leaks within the system and can be used as a means of leak detection. To be practical, some examples can be given; those that use: (i) inverse methods to determine parameters in transient models by comparison with observed data (inverse transient analysis), (ii) transient damping-free-vibrational analysis, and also (iii) methods that use the time of arrival and magnitude of leak-reflected signals in order to determine leak location. The

Article history: received on Mar. 11, 2016; accepted on Aug. 23, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0049.

subsequent pressure response due to leakage is measured and analysed to derive useful system information. The above approach is generally known as system response extraction and forms the main aspects of well-established methodologies. They are used to extract dynamic responses [2] of complex mechanical and electrical systems. Diverse spectral techniques, more robust than STFT and FFT have been implemented. Among them, we recall FDM (*Filter Diagonalization Method*), DSD (*Decimated Signal Diagonalization*) and DPA (*Decimated Padé Approximant*). FDM is a robust technique capable of avoiding the influence of noise for huge signals to be processed [3, 4]. DSD is still robust but fast and more accurate than FDM for a portion of signals to be treated [5] while DPA is a specific case of processing able to implement fast approximation with a powerful quickness [6]. All three techniques are suitable to convert nonlinearities into linear items, exploiting the linear algebra.

For small transient signals, the impact of nonlinearity in pipeline systems is negligible and, for these cases, a pipeline can be considered as a linear system. Whereas resonance frequencies reinforce and transmit input signals, other frequencies are absorbed within the system [7]. In this respect, pipeline systems are similar to frequency filters, the characteristics of which are determined by system properties such as boundary conditions, friction, and wave speed. Using the convolution theorem regarding the Fourier Transform, it is possible to define an inverse signal in respect to convolution. Suppose that we have a signal g and we would like to find a signal (if it exists) g^{-1} with the property that $g * g^{-1} = \delta$, in which " * " is intended as a cross product. The transform of a δ function (Kronecker in discrete time and Dirac in continuous time) is just a constant 1.

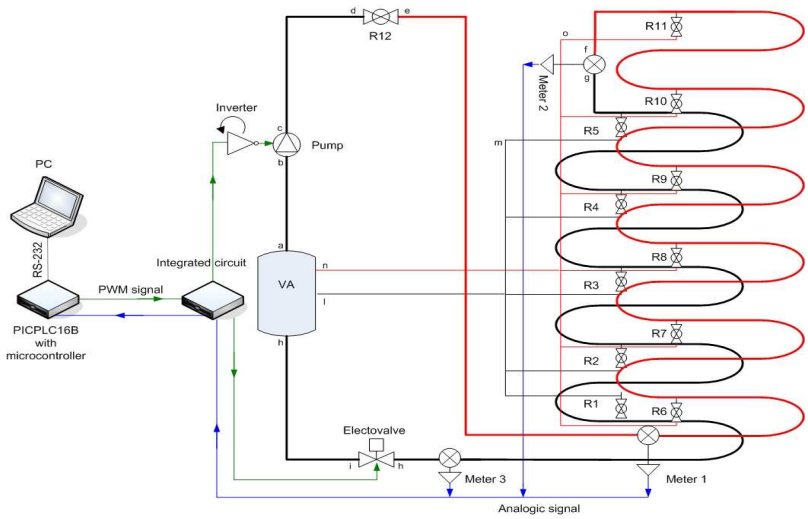

Fig. 1. Hydraulic and acquisition architectures of a pipeline hung on the lab wall.

The created hydraulic system uses a special zigzag pipeline simulating an experimental and complex waterworks 120 meters long and with a diameter of 1 inch. It is made of copper coated with a plastic film. It contains 11 water taps located on the pipeline and 3 magnetic sensors. A dedicated electronic architecture has been implemented for acquiring and processing signals produced by the magnetic sensors. A scheme of the architecture is shown in Fig. 1, whereas its photo – in Fig. 2. The system is hung on a vertical wall of the laboratory of Measurement and Instrumentation of the Department of Innovation Engineering – University of Salento, Italy. The fundamental interest, as announced before, is in trying to use the STFT and FFT techniques,

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 631–644.

which – even if not robust for the purpose of the research can be employed thanks to some improvements, *e.g.* a further demonstrated *ad hoc* solution to ill-posed problems using an *L*-curve approach.

The experimental activities have been carried out starting with the regulation of water taps (see Fig. 3) to understand the effects on the water trends within the pipeline. A reservoir shown in Fig. 2 serves for storing approximately 100 litres of water used for filling the pipeline.

Fig. 2. A photo of the experimental system hung on a lab wall.

Actually, the pipeline has two parts, the first being an old circuit constructed with 6 water taps and one magnetic transducer. To make it more complex, a second part was added including 5 new water taps (Fig. 3), and it was superimposed to the first part. This is a stressing configuration, very complex for experimental activities. Both previous and new portions have a certain influence on leak position recovery as will be seen in the results' section.

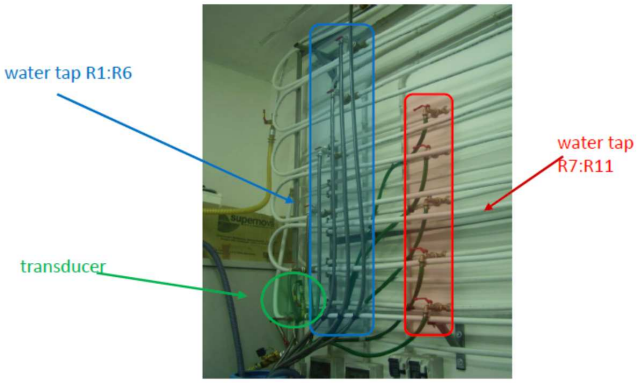

Fig. 3. A zoomed view of all water taps simulating leaks.

Even so we have experimented with other techniques on the same hydraulic circuit, we report the calibration that leads to a correlation between peak and opening/closing manoeuvers of water taps. That is an essential procedure on the proviso that the right peak should be found. The valve or water tap status and its duration enable to determine the trend at a certain pressure produced by the pump. Pressure fluctuations are studied for any water tap as shown in Fig. 4, in order to understand the eventual clutter pressure. For the cases under test, the pump delivers water at the same pressure in order to simulate a real waterworks. That is carried out thanks to a specially implemented electronic control.

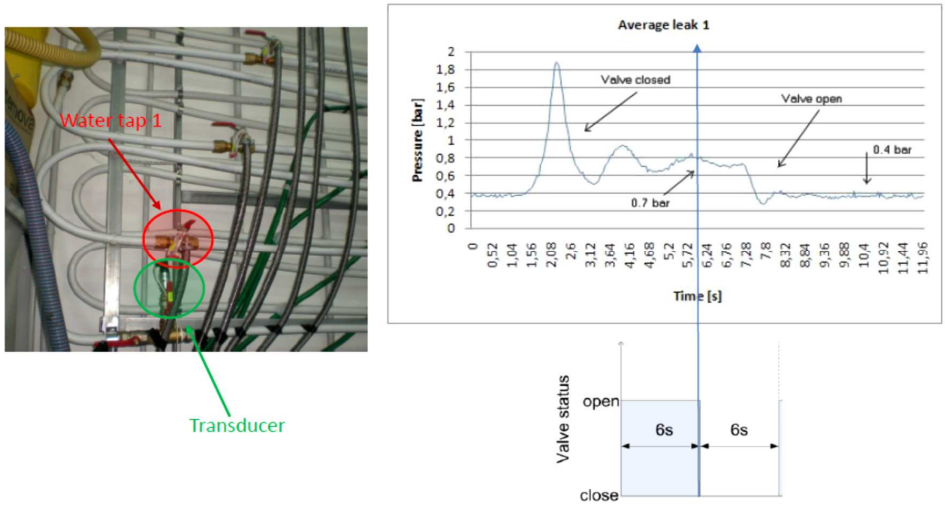

Fig. 4. Trials for calibrating the system and for detecting leaks by opening and closing a water tap.

2. STFT approach

The STFT represents a sort of compromise between the time- and frequency-based views of a signal. It provides some information about both when and at what frequencies a signal event occurs. However, we can only obtain this information with a limited precision and such a precision is determined by the size of the window. As described in Section 1, vibrations are created during leak, and using an electronic instrument or a proper sensing device it is possible to convert them in sounds or – equivalently – in the sum of sinusoids. So, the spectral responses are characterized by spectrograms that are produced by a procedure known as the *Short-Time Fourier Transform* (STFT). The STFT divides the entire signal into a series of successive short-time segments, called *records* (or *frames*). Each record is used as an input to the *Discrete Fourier Transform* (DFT), generating a series of spectra (one for each record).

Let $(X(t))_{t\in Z}$ be a digital signal. We review here the conditions for perfect reconstruction of the signal through STFT and inverse STFT [8]. Let N be a window length, R a window shift, W an analysis window function and S a synthesis window function. We assume that W and S are zero outside an interval $0 \le t \le N-1$. Also, we assume that the window length N is an integer multiple of the shift R and we note $Q = N/R$. The STFT for frame m is defined as the DFT of the windowed short-time signal $W(t - mR)\, X(t)$ (with the phase origin at the start of the frame, $t = mR$). The inverse STFT procedure consists in Fourier-inverting of each frame of the STFT spectrogram, multiplying each obtained (periodic) short-time signal by a synthesis window and summing together all the windowed short-time signals. In a particular frame $mR \le t \le mR+N-1$, which leads to a reconstructed signal $Y(t)$ given by:

$$\begin{aligned} Y(t) &= S(t - mR)\, W(t - mR)\, X(t) \\ &+ \sum_{q=1}^{Q-1} S(t - m(m-q)R)\, W(t - (m-q)R)\, X(t) \\ &+ \sum_{q=1}^{Q-1} S(t - m(m+q)R)\, W(t - (m+q)R)\, X(t)\,, \end{aligned} \tag{1}$$

where the three terms on the right-hand side are respectively a contribution of the inverse transforms of frame m, overlapping frames on the left and overlapping frames on the right. As

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 631–644.

the contributions of frames with an index difference larger than Q do not overlap, by equating $Y(t) = X(t)$ for all t, we obtain as in [9] the following necessary condition for perfect reconstruction:

$$1 = \sum_{q=0}^{Q-1} W(t - qR)S(t - qR). \tag{2}$$

3. Implemented FFT and STFT algorithms for spectral analysis

With $|\cdot)$ we denote a representation of vectors in a vector space H; given two vectors $|a), |b) \in H$ the notation $(\cdot|\cdot)$ indicates a complex symmetric inner product such as $(a|b) = (b|a)$. During computation of this product, we only carried out transposition but not the complex conjugate. For example, given the following vectors: $|a) = \begin{bmatrix} j & 1-j \end{bmatrix}^T$ and $|b) = \begin{bmatrix} 2j & 1 \end{bmatrix}^T$ we have:

$$(a|b) = (b|a) = \begin{bmatrix} j & 1-j \end{bmatrix} \begin{bmatrix} 2j \\ 1 \end{bmatrix} = \begin{bmatrix} 2j & 1 \end{bmatrix} \begin{bmatrix} j \\ 1-j \end{bmatrix} = -1 - j. \tag{3}$$

We identify a linear operator on the vector space H with a superscript $\hat{}$, *e.g.* $\hat{U}, \hat{\Omega}$, *etc.* To indicate the application of an operator to a vector, we denote $|b) = \hat{\Omega}|a)$. An operator is defined diagonalisable if it is itself a set of eigenvalues ω_k and eigenvectors $|\omega_k)$ such as :

$$\hat{\Omega}|\omega_k) = \omega_k|\omega_k), \tag{4}$$

where the eigenvectors are orthonormalized in respect to the complex symmetric inner product:

$$(\omega_k|\omega_{k'}) = \delta_{kk'}. \tag{5}$$

When the eigenvectors $|\omega_k)$ constitute a complete basis, for the operator identity it is:

$$\hat{I} = \sum_k |\omega_k)(\omega_k|. \tag{6}$$

That implies that we can write $\hat{\Omega}$ by means of its spectral representation:

$$\hat{\Omega} = \sum_k \omega_k |\omega_k)(\omega_k|. \tag{7}$$

A spectral representation is useful when it is necessary to obtain information from a spectral function $f(\hat{\Omega})$ of operator $\hat{\Omega}$ for which eigenvalues and eigenvectors are known:

$$f(\hat{\Omega}) = \sum_k f(\omega_k)|\omega_k)(\omega_k|. \tag{8}$$

The function $f(\hat{\Omega})$ is also an operator, with eigenvalues $f(\omega_k)$ and eigenvectors $|\omega_k)$. In terms of quantum mechanics, since we deal with vibrations within the pipelines, if $\hat{\Omega}$ identifies an operator which is linear, hamiltonian and symmetric, with eigenvalues ω_k and eigenvectors $|\omega_k)$, it will be of great importance to use the associated temporal evolution operator: $\hat{U} = e^{-j\hat{\Omega}t}$. In fact, if $|0)$ is the system initial state, a state $|t)$ at a moment t is given by:

$$|t) = \hat{U}(t)|0). \tag{9}$$

The time-dependent autocorrelation functionis then given by:

$$\xi(t) = (0 \mid t) = (0 \mid \hat{U}(t) \mid 0) = (0 \mid e^{-j\hat{\Omega}t} \mid 0). \tag{10}$$

According to (8),the spectral representation $\hat{U}$ becomes:

$$\hat{U}(t) = f(\hat{\Omega}) = e^{-it\hat{\Omega}} = \sum_k e^{-it\omega_k} \mid \omega_k)(\omega_k \mid . \tag{11}$$

To show how the algorithm works, it is necessary to modify the FFT algorithm by defining a complex one-dimensional signal in the time domain, $c_n = C(t_n)$, defined in a set of equidistant time intervals $t_n = n\tau$, $n = 0,1,...,N-1$ as the sum of damped sinusoids:

$$c_n = \sum_{k=1}^{K} d_k e^{-in\tau\omega_k} = \sum_{k=1}^{K} d_k e^{-in\tau(2\pi f_k - i\gamma_k)} \tag{12}$$

with a total of $2K$ unknowns, that are K complex amplitudes d_s and K complex frequencies $\omega_k = 2\pi f_k - i\gamma_k$ that also include damping. Although (4) is nonlinear, its solution can be obtained with linear algebraic methods. The proposed FFT [12] associates an autocorrelation function, in an appropriate dynamic time system described by a complex Hamiltonian operator $\hat{\Omega}$ with complex eigenvalues $\{\omega_k\}$, with a signal c_n to be transformed in the form of (8):

$$c_n = \left(\Phi_0 \middle| e^{-in\tau\hat{\Omega}} \Phi_0\right). \tag{13}$$

In this way, the problem can be reduced to diagonalization of the Hamiltonian operator $\hat{\Omega}$ or, similarly, the evolution operator $\hat{U} = \exp(-i\tau\Omega)$.

A complex inner symmetric product operation is used in (13), namely $(a|b) = (b|a)$ without a complex conjugation, and Φ_0 is the initial state. Again, the symbol (.|.) denotes a complex symmetric inner product. Assuming we have a set of orthonormal eigenvectors $\{Y_k\}$ that diagonalize the evolution operator, we can clarify it as:

$$\hat{U} = \sum_k u_k \left|Y_k\right)(Y_k\right| = \sum_k \exp(-i\omega_k\tau) \left|Y_k\right)(Y_k\right| \tag{14}$$

and substituting (14) in (13), remembering to let:

$$d_k = (\Phi_0 \left|Y_k\right)(Y_k \left|\psi_0\right) = (Y_k \left|\Phi_0\right)^2 . \tag{15}$$

The computed eigenvalues determine the positions of spectrum lines and their widths while the eigenvectors define their amplitudes and phases. Let us adopt a simple set created from Krylov vectors [11], generated by the evolution operator: $\Phi_n = \hat{U}^n \Phi_0 = \exp(-in\tau\hat{\Omega})\Phi_0$. According to (7), it gives:

$$(\Phi_n \left|\hat{U}\Phi_m\right) = (\Phi_n \left|\Phi_{m+1}\right) = c_{m+n+1}\,, \tag{16}$$

but since the set is not orthonormal, and we define a subspace of Krylov vectors generated by vectors of $\boldsymbol{Q}$, *i.e.* $Q = \{|0),|1),\ldots,|M-1)\}$, the overlap matrix should be computed as follows:

$$(\Phi_n \left|\Phi_m\right) = (\hat{U}^n \Phi_0 \left|\hat{U}^m \Phi_0\right) = (\Phi_0 \left|\hat{U}^{m+n} \Phi_0\right) = c_{m+n+1} . \tag{17}$$

Therefore, it is strictly related to the values of measured signal. Then the following notation could be used: $\boldsymbol{U}^0$ is a representation of the $(M+1)$ *x* $(M+1)$ overlap matrix, similarly $\boldsymbol{U}^1$ is for $\hat{U}$. To signalize the formulation of (12), one must solve the generalized problem of eigenvalues, *i.e.*:

$$\mathbf{U}^1\mathbf{B}_k = u_k \mathbf{U}^0 \mathbf{B}_k\,, \tag{18}$$

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 631–644.

in which $u_k = \exp(-in\omega_k\tau)$ gives lines of spectrum and their widths, whereas eigenvectors $\boldsymbol{B}_k$ give amplitudes and phases. The matrix $\mathbf{B}_k$ is derived from the below considerations. Let us assume that the generic eigenvector $|\,\omega_k)$ can be expressed as a linear combination of elements of $\boldsymbol{Q}$. We define:

$$\mathbf{V} = \left[|\,0) \quad |\,1) \quad \ldots \quad |\,M-1)\right] = \left[|\,0) \quad \hat{U}\,|\,0) \quad \ldots \quad \hat{U}^{M-1}\,|\,0)\right] \in \mathbb{C}^{M,M}, \tag{19}$$

as a matrix having the vectors of base $\boldsymbol{Q}$ as its columns, and:

$$\mathbf{B}_k = \begin{bmatrix} B_{k,0} \\ B_{k,1} \\ \\ B_{k,M-1} \end{bmatrix} \in \mathbb{C}^{M,1}, \tag{20}$$

as a vector of appropriate coefficients related to the eigenvectors $|\,\omega_k)$. We can write:

$$|\,\omega_k) = \left[|\,0) \quad |\,1) \quad \ldots \quad |\,M-1)\right] \begin{bmatrix} B_{k,0} \\ B_{k,1} \\ \vdots \\ B_{k,M-1} \end{bmatrix} = \mathbf{V}\mathbf{B}_k. \tag{21}$$

By substituting (21) in the implicit form of (4) the following is obtained:

$$\hat{U}\mathbf{V}\mathbf{B}_k = u_k\mathbf{V}\mathbf{B}_k. \tag{22}$$

By further implicit considerations, pre-multiplying both members by $\mathbf{V}^T$, we obtain:

$$\mathbf{V}^T\hat{U}\mathbf{V}\mathbf{B}_k = \mathbf{V}^T u_k\mathbf{V}\mathbf{B}_k. \tag{23}$$

That leads to (18) which is the generalized eigenvalues problem, where u_k are eigenvalues and $\mathbf{B}_k$ eigenvectors. After solving the problem, having calculated u_k and $\mathbf{B}_k$, the frequencies ω_k are determined by using the following formula:

$$\omega_k = -\frac{1}{\tau}\measuredangle(u_k) = -\frac{1}{\tau}\measuredangle(e^{-j\tau\omega_k}), \tag{24}$$

where τ is a sampling time and a symbol $\measuredangle$ applied to a complex number delivers its phase, *i.e.* $\measuredangle e^{j\vartheta} = \vartheta$. To determine the amplitudes, we employ:

$$d_k = (0\,|\,\omega_k)^2, \tag{25}$$

or, since $|\,\omega_k) = \mathbf{V}\mathbf{B}_k$, after pre-multiplying both members by the vector row $(0\,|$, the previous equation can be also expressed as:

$$(0\,|\,\omega_k) = (0\,|\,\mathbf{V}\mathbf{B}_k. \tag{26}$$

Further on, we obtain:

$$(0\,|\,\omega_k) = (0\,|\left[|\,0) \quad |\,1) \quad \ldots \quad |\,M-1)\right]\mathbf{B}_k. \tag{27}$$

In the aforementioned considerations we have assumed $\mathbf{U}^1, \mathbf{U}^0 \in \mathbb{C}^{M,M}$ symmetric. Based on the previous explanations, now the algorithms should be clear. The software interpreting the algorithm, *i.e.* FFT or STFT, starts with a specific icon where the operator must include essential parameters to start with the acquisition according to (12). The icon is presented in Fig. 5, along with STFT processing. The following parameters are set: type of window (rectangular), sampling rate, window length, step and padding. Flowcharts of FFT and STFT algorithms are shown in Fig. 6.

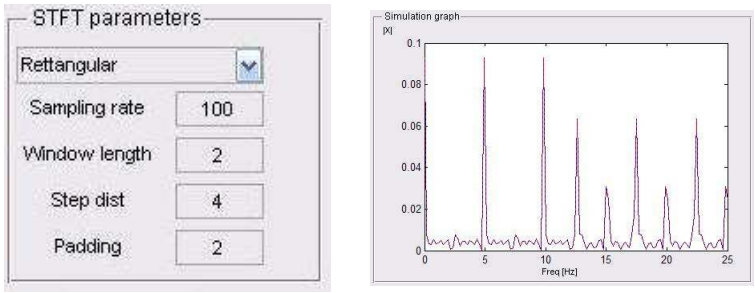

Fig. 5. The set parameters (left) and processed signal according to STFT (right).

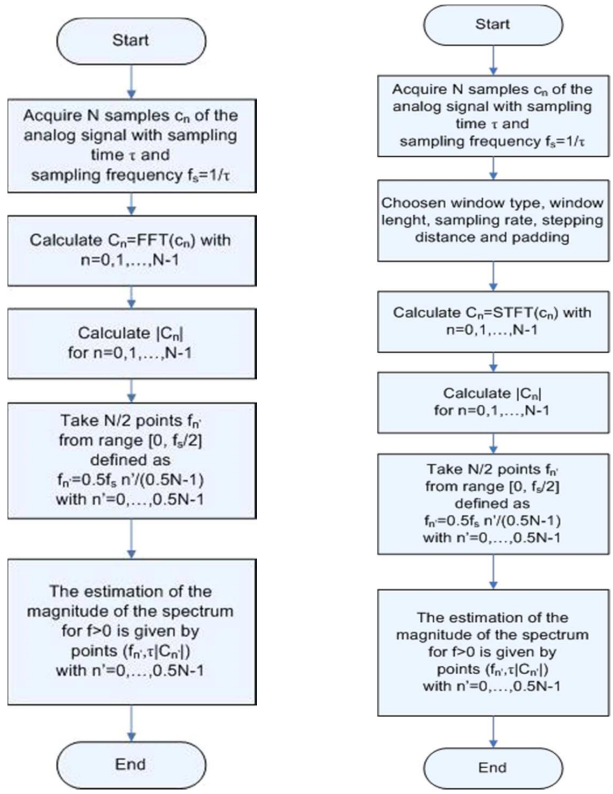

Fig. 6. Flowcharts of the implemented FFT (on the left) and STFT (on the right) algorithms for leak detection.

However, as it will be seen in the results' section, due to intrinsic limitations of FFT and much more those of STFT, we do not obtain better results than we do with FDM, DSD and PDA. That is related to uncertainty values obtained (see 18) with the use of eigenvalues and eigenvectors that are greater than those obtained with other robust techniques. To overcome this key disadvantage, we have implemented, taking inspiration from the Tikhonov regularization method [12], a dedicated algorithm based on an *L*-curve approach [13].

Let us consider a matrix A in the following linear equation:

$$b = Ax + w\,, \tag{28}$$

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 631–644.

where: b is a vector of observation; A is a matrix that describes distortion caused by the system under test; x is an unknown object of interest and w is a random vector that represents additional noise. We state that this discrete problem is ill-posed if the following conditions are met:

1. Single values of A gradually decline to zero.
2. The ratio of the greatest single value and the smallest (not null) one is great.

The first condition indicates that in the vicinity there are no problems with a matrix of good-posed coefficients and a well determined numerical rank. The second criterion implies that the matrix is ill-posed, i.e. the solution is sensitive to perturbations. In many cases it happens that matrix A is ill-posed and the main difficulty in ill-posed issues is that they are essentially undetermined because of small single values of A. In the effort to stabilize the problem, adding further information to the desired solution is required: that is the regularization method. It typically requires that a norm 2 of the solution must be small. It is also possible to include an estimation of the solution x_0 in the constraint. The constraint is:

$$\min \Omega(x) \qquad with \;\; \Omega(x) = \left\| L(x - x_0) \right\|. \tag{29}$$

The matrix L can be:

a) Typically, an identity matrix I_n;

b) A discrete approximation $P \times N$ of the derivation operator (n-p) *i-th*.

We define a regularized solution x_q that can minimize the following weighted one of the combination of the residual norm and constraint:

$$x_q = \min_x \left\{ \left\| Ax - b \right\|^2 + q^2 \left\| Lx - x_0 \right\|^2 \right\}, \tag{30}$$

in which $q > 0$ is a regularization parameter:

- for a great q (a great quantity of regularization) a solution agrees with a small norm at the cost of a great residual norm;
- a small q (a small quantity of regularization) has the opposite effect.

Equation 13 can be generalized in the following way:

$$x_q = \arg\min_x \Phi(b, x) + q\Psi(x). \tag{31}$$

A graphical tool more convenient for analysis of discrete ill-posed problems is the so-called L-curve, that means a plot of norm $\left\| Lx_q \right\|$ of a regularized solution in respect to the residual norm $\left\| Ax_q - b \right\|$. In this way, the L-curve clearly demonstrates a compromise between the minimization of both quantities. When q is too great (over-regularization), the curve is essentially a horizontal line. Vice versa, when q is too small (under-regularization), the curve is mainly a vertical line according to Fig. 7 (on the left); q displays a characteristic shape of L. The transition between these two regions, over and under regularization, corresponds to the angle of L-curve, and its relative value of q at this angle (called L-corner) is proposed as the optimum value for q. A flowchart on the right of Fig. 7 shows the algorithm we have implemented to overcome the limitations of using FFT and STFT we recalled before. The algorithm also includes a block of SVD (*Single Value Decomposition*) [14].

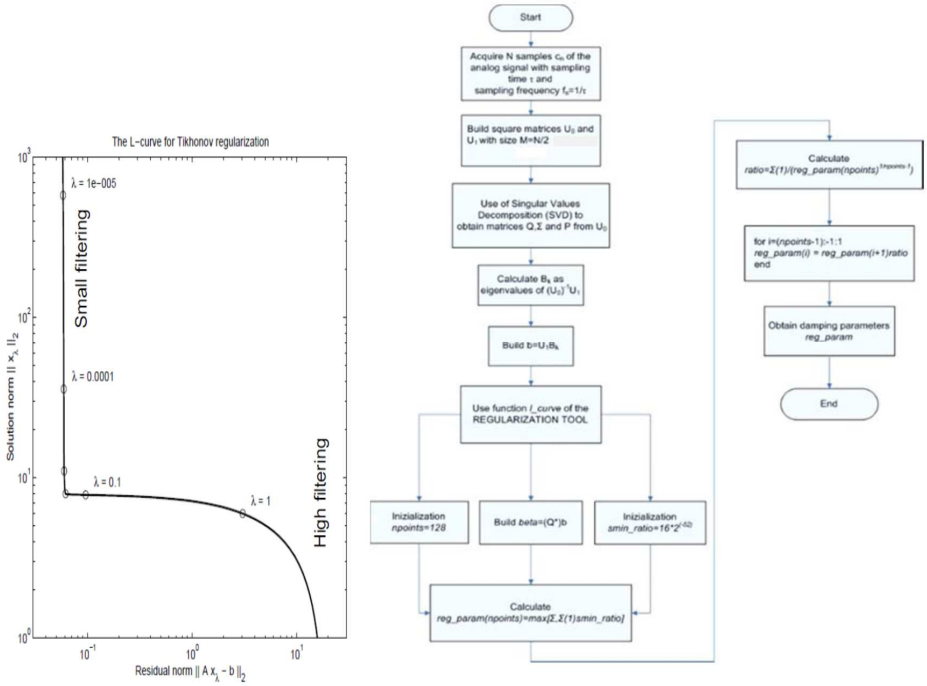

Fig. 7. An *L*-curve for improving the Tikhonov regularization (left) and a flowchart of the proposed algorithm (right).

4. Results and discussion

Before obtaining the values of amplitude, we have performed 5 cycles of tests per water tap, *i.e.* 5 tests for detecting the *j*-th leak with $j = 1,....,11$; an average of all 5 acquired waveforms is as follows:

$$p_j(t) = \frac{1}{5}\sum_{i=1}^{5} p_{i,j}(t)\,. \tag{32}$$

In this way, the obtained signal preserves significant characteristics while the noise is included in the measurements $p_{i,j}(t)$. The explanation of this technical attitude depends upon the fact according to which:

$$p_{i,j}(t) = \overline{p}_{i,j}(t) + n_{i,j}(t), \tag{33}$$

with $\overline{p}_{i,j}(t)$ as a true value of pressure and noise of measurement $n_{i,j}(t)$. The quantity evaluated in $t = t^*$ can be modelled, when i varies, as an aleatory variable with zero average. By considering the 5 variables $n_{i,j}(t^*)$ $i = 1, \ldots, 5$, we obtain:

$$E[n_{i,j}(t^{"})] = \frac{1}{5}\sum_{i=1}^{5} n_{i,j}(t^{"}) \simeq 0\,. \tag{34}$$

The above procedure is repeated for each water tap (valve). At the conclusion of measurements, we have 11 waveforms $p_j(t)$, $j = 1,...,11$, each one describing the behaviour of the system for a given condition of leakage.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 631–644.

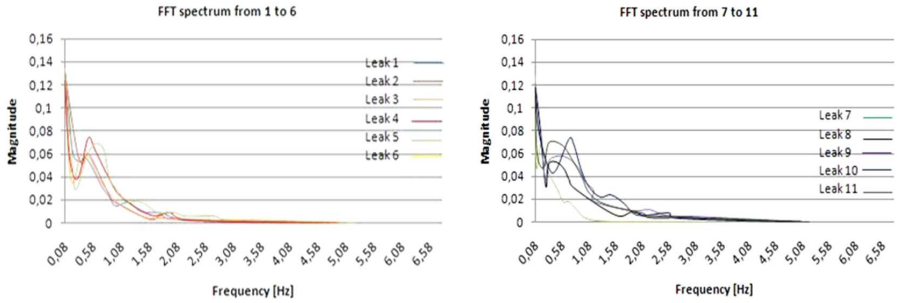

Fig. 8. Different leaks recovered by FFT after single opening of 11 water taps.

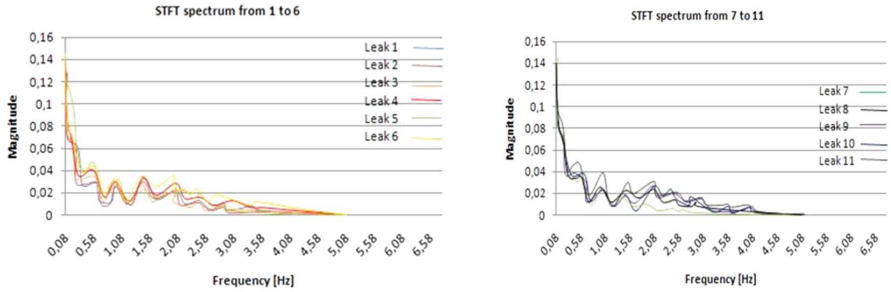

Fig. 9. Different leaks recovered by STFT after single opening of 11 water taps.

Now, we can comment on the results of the previous procedures by applying the algorithms of Fig. 6 that bring the waveforms of Fig. 8 and Fig. 9 for FFT and STFT, respectively. We can see the intrinsic behaviour of both algorithms in the same conditions. Given for instance 1.08 Hz, as we should expect, FFT displays the major peak greater than that of STFT; the same takes place for all useful frequencies. As stated before, we encounter the ill-posed problems that mostly influence the determination of leak locations. The results of Fig. 8 and Fig. 9 must be further treated since the positions of eigenvalues can be located on a circle of radius 1 as depicted in Fig. 10. So our goal is to overcome the ill-posed issue by implementing the algorithm from Fig. 7.

However, it is necessary to notice that the uncertainty is obtained using a specific method for this scope. The data recovered after acquisitions are used for the determination of uncertainty by means of the least mean squares/linear regression. So we should start with the calculation of coefficients of a regression straight line:

$$y = ax + b\,, \tag{35}$$

so that a distance between points is minimal. In (35), we denote x_i expressed in *metres*, and it represents the distance at which we encounter the leak i, with $i = 1,\ldots,11$, measured from the pressure transducer, whilst y_i is the peak height within the spectrum of Figs. 8 and 9. To retrieve constants of (35), the following formulae are used:

$$b = \frac{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{\Delta}\,,$$

$$a = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\Delta}\,, \tag{36}$$

$$\Delta = N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2,$$

$$\sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - b - ax_i)^2}\,.$$

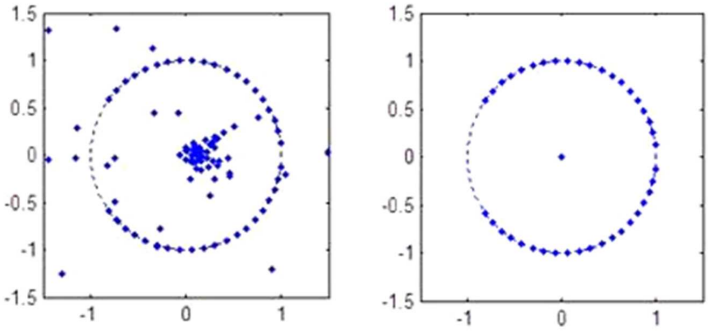

Fig. 10. Locations of eigenvalues before (on the left), and after (on the right) the *L*-curve implementation.

in which a and b are from (35), Δ is a deviation, and σ_y is an uncertainty of amplitude; an uncertainty of distance σ_x is given by the variable x value obtained by reversing (35), *i.e.* $x = \dfrac{y-b}{a}$, and calculating the uncertainty as:

$$\sigma_x = \left[\frac{dx}{dy}\right]\sigma_y = \frac{1}{a}\sigma_y\,. \tag{37}$$

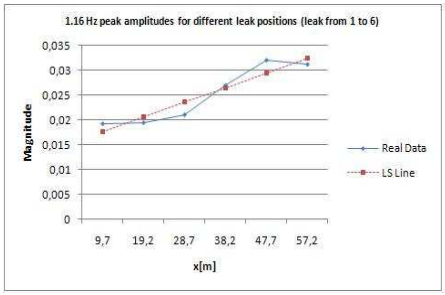

Fig. 11. Interpolation based on the FFT signal of points with peak heights at f =1 Hz taking into account the leak position.

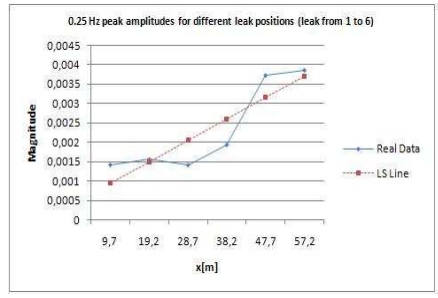

Fig. 12 Interpolation based on the STFT signal of points with peak heights at f = 0.25 Hz taking into account the leak position.

The application of linear regression either FFT or STFT is displayed in Fig. 11 and Fig. 12, at least for the first portion of leaks; LS stands for the least squares method. The plots interpolate the experimental points in a sense of least squares. The final results are shown in Table 1 for FFT and STFT, respectively. For each technique, we present a double result. The first result is related to the use of an *L*-curve that leads to all eigenvalues located on the edge of the circle. The second, instead, reports the implementation of the algorithm without an *L*-curve. As a matter of fact, an *L*-curve gives a great opportunity to reduce the uncertainty. The uncertainty demonstrates the position of leak in respect to the sensor location. It is intuitive to understand that as we move farther from the transducer, the detection of leak location becomes less precise and the uncertainty of it increases.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 631–644.

Table 1. The summarized results of FFT and STFT algorithms.
The improved algorithms are related to eigenvalues located on a circle.

Water tap	Technique	Eigenvalues position	Uncertainty
*R*1 : *R*6	FFT	On circle	±5.20 m
*R*1 : *R*6	FFT	Not fully on circle	±11.52 m
*R*1 : *R*6	STFT	On circle	±7.92 m
*R*1 : *R*6	STFT	Not fully on circle	±12.40 m
*R*7 : *R* 11	FFT	On circle	±5.22 m
*R*7 : *R* 11	FFT	Not fully on circle	±6.90 m
*R*7 : *R* 11	STFT	On circle	±2.60 m
*R*7 : *R* 11	STFT	Not fully on circle	±3.42 m

Certainly, the above results are "worse" in respect with those attained by means of DSD and FDM as reported in references. However, for normal but not complicated waterworks, where we do not generally deal with sudden pressure variations and huge piezo-metric heights, the proposed approach is reliable. For these configurations, the approach is not time-consuming and can offer similar and comparable results.

5. Conclusions

We have presented an enhancement of FFT and STFT techniques for leak detection by applying the *L*-curve approach in accordance with the Tikhonov technique. The Tikhonov regularization is much less numerically expensive than other regularization techniques (as SVD) and reaches its aim of removing the singularity in the denominator, because the new matrix is a Hermitian and positive definite one. FFT is, by definition, the golden standard method of spectral analysis. But, for complex architectures, it displays limitations as well as STFT does. Table I shows comparison of the two algorithms; the improvements are noticed with the implementation of a regularization technique based on an enhanced Tikhonov technique. In general, the FFT algorithm offers better results than the STFT one; but in some circumstances, for specific conditions, STFT can display better results in comparison with FFT. In general, the method of FFT does not provide simple global results for Fourier representations of the input. Such generality and simplicity are usually possible only for linear systems, as for the experimental zig-zag plant of this paper. General results can be obtained for memoryless nonlinearities operating on sinusoidal inputs. This result is not as important as the corresponding result for linear systems because sinusoids are not fundamental building blocks of nonlinear systems, in contrary to the linear ones. This research shows [15] that it is possible to estimate the detection of leaks with good accuracy for zigzag pipelines that have a diameter of less than 20 cm, and we can arrive to around 1 inch as it is done in this paper. These results open opportunities for implementing the algorithm for pipelines used to constitute *e.g.* industrial heat exchangers, or to improve reliability of normal pipelines [16].

References

[1] Kapelan, Z., Savic, D., Walters, G., Covas, D., Graham, N., Maksimovic, C. (2003). An Assessment of the Application of Inverse Transient Analysis for Leak Detection: Part I – Theoretical Considerations. *Computer Control for Water Industry*, London, UK.

[2] Lee, P., Vítkovský, J., Mohapatra, P.K., Chaudhry, M.H., Kassem, A.A., Moloo, J. (2006). Detection of Partial Blockage in Single Pipelines. *Journal of Hydraulic Engineering*, ASCE, 132(2), 200–206.

[3] Lay-Ekuakille, A., Vendramin, G., Trotta, A. (2009). Spectral Analysis of Leak Detection in a Zigzag Pipeline: A Filter Diagonalization Method – based algorithm application. *Measurement*, 42(3), 358–367.

[4] Lay-Ekuakille, A., Vendramin, G., Trotta, A. (2010). Robust Spectral Leak Detection of Complex Pipelines using Filter Diagonalization Method. *IEEE Sensors Journal*, 9(11), 1605–1614.

[5] Lay-Ekuakille, A., Vergallo, P. (2014). Decimated Signal Diagonalization Method for Improved Spectral Leak Detection in Pipelines. *IEEE Sensors Journal*, 14(6), 1741–1748.

[6] Lay-Ekuakille, A., Vergallo, P., Griffo, G. (2013). A Robust Algorithm based on Decimated Padé Approximant Technique for Processing Sensor Data in Leak detection in Waterworks. *IET Science, Measurement & Technology*, 7(5), 256–264.

[7] Griffin, D.W., Lim, J.S. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoustics, Speech, and Signal Proc.*, 32(2), 236–243.

[8] Lay-Ekuakille, A., Vendramin, G,, Trotta, A., Vanderbemdem, P. (2009). STFT-based spectral analysis of urban waterworks leakage detection. *XIX IMEKO World Congress Proc.*, Lisbon, Portugal.

[9] Liou, C.P., Tian, J. (1995). Leak Detection – A Transient Flow Simulation Approach. *Journal of Energy Resources Technology, American Society of Mechanical Engineers*, 117(3), 243–248.

[10] Nash, G.A., Karney B.W. (1999), Efficient Inverse Transient Analysis in Series Pipe Systems. J*ournal of Hydraulic Engineering*, 125(7), 761–764.

[11] Lay-Ekuakille, A., Pariset, C., Trotta, A. (2010). FDM-based Leak Detection of Complex Pipelines: Robust Technique for Eigenvalues Assessment. *Measurements Science Technology*, 21, 1–10.

[12] Walkins, D.S. (2007). *The matrix Eigen Problem*. SIAM, 351–421.

[13] Calvetti, D., *et al.* (2000). Tikhonov regularization and the L-curve for large discrete ill-posed problems. *Journal of Computational and Applied Mathematics*, 123(1–2), 423–446.

[14] Hansen, P.C. (1998). *Rank-Deficient and Discrete Ill-posed Problems*. Siam, Philadelphia, USA.

[15] Lay-Ekuakille, A., Vergallo, P., Trotta, A. (2010). Impedance Method for Urban Waterworks: Experimental Frequency Analysis for Leakage Detection. *Imeko Tc-4, TC-19 and IWADC Conference*, Kosice, Slovakia.

[16] Ozevin, D., Yalcinkaya, H. (2014). New Leak Localization Approach in Pipelines Using Single-Point Measurement. *Journal of Pipeline Systems Engineering and Practice*, 5(2), 1–8.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 645–652.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

MEASUREMENT OF NOISE IN SUPERCAPACITORS

Arkadiusz Szewczyk

Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics, G. Narutowicza 11/12, 80-233 Gdańsk, Poland
(✉ szewczyk@eti.pg.edu.pl, +48 58 347 2140)

Abstract

A developed method and measurement setup for measurement of noise generated in a supercapacitor is presented. The requirements for noise data recording are considered and correlated with working modes of supercapacitors. An example of results of low-frequency noise measurements in commercially available supercapacitors are presented. The ability of flicker noise measurements suggests that they can be used to assess quality of tested supercapacitors.

Keywords: supercapacitor, flicker noise, measurement set-up, reliability.

1. Introduction

Noise is known as an indicator for assessing quality and reliability of devices. It is widely used for semiconductor devices, sensors of various characters, electrochemical units, chemical reactions as corrosion, and other random phenomena [1−5]. Also, the use of noise methods can be studied for the assessment of capacitors' quality [6]. A detailed procedure is not obvious, as those devices are commonly used as elements for suppressing noise from circuits and therefore $1/f$ noise can be dominant at a very low frequency range only.

A supercapacitor is an electronic device that is capable of storing a relatively high amount of energy in comparison with its mass. On a Ragone plot, supercapacitors are placed between electrolytic capacitors and batteries [7]. Thanks to a very low series resistance, a supercapacitor can be charged and discharged very fast with a relatively high current. This, combined with a high number of charging-discharging cycles predestines it for applications that require management of peak powers and high dynamics. Typical applications of supercapacitors are energy storage systems with high current peaks, as in automotive applications for energy retrieval, energy harvesting and combined battery-supercapacitor systems.

Increasing popularity of supercapacitors and growing market of this devices require continuous development of methods for assessment of their quality and reliability. Nowadays, the most popular and commonly used methods for testing supercapacitors are: *cycling voltammetry* (CV), *galvano-static cycling with potential limitations* (GCPL), impedance spectroscopy and accelerated aging [8, 9]. All those methods are based on the observation of current or voltage during forced charging/discharging of a supercapacitor in various voltage and current conditions.

Quality of a supercapacitor is usually derived from its capacitance, equivalent series resistance, ESR, and impedance. Degradation of supercapacitor is indicated by the change of its capacity and ESR and is measured by known methods of estimation of those parameters.

Article history: received on Apr. 28, 2017; accepted on Jul. 31, 2017; available online on Dec. 10, 2017; DOI: 10.1515/mms-2017-0059.

2. Equivalent circuit of supercapacitor

One of types of supercapacitors is an *electric double layer capacitor* (EDLC). The EDLC comprises two porous carbon electrodes with an ion permeable separator and electrolyte solution between them. A typical EDLC structure is shown in Fig. 1.

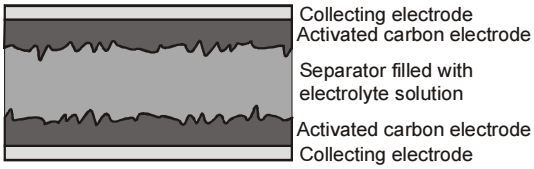

Fig. 1. An illustration of supercapacitor's structure.

When a supercapacitor is cyclically charged and discharged, four stages can be distinguished: 1) charging, when charges flow into the structure and a voltage increase is observed at the capacitor terminals; 2) a voltage drop, after the capacitor is charged and left with open terminals (the voltage at the terminals slowly decreases); 3) discharging, when charges flow out of the structure of capacitor, and a voltage drop is observed at the terminals and 4) the voltage restore, when the capacitor is discharged and left open-circuit (a voltage increase is observed between the terminals). A voltage curve when charging the supercapacitor with a constant current at disconnected terminals and discharging it with a constant current is shown in Fig. 2.

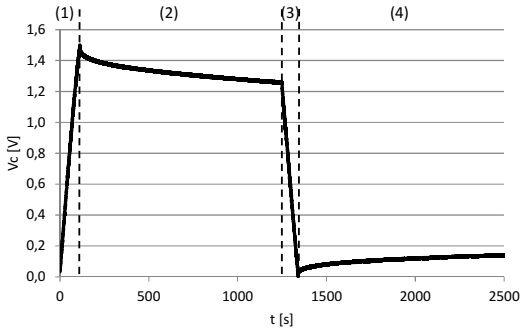

Fig. 2. A change of voltage between supercapacitor terminals during charging (1), when disconnected (2), (4) and during discharging (3).

An electrical equivalent circuit of supercapacitor that models its behaviour during the charging – discharging process is described in [10, 11]. The model comprises two branches, as shown in Fig. 3. The first branch with the equivalent series resistance ESR and capacitance C_H represents the Helmholtz layer capacity available for fast charging/discharging. The second one, with capacitance C_D and resistance R_D represents the mechanism of charges' redistribution by the diffusion mechanism [10, 12]. The electric capacitance of diffusion mechanism is represented by the capacitor C_D. The resistance R_D determines how fast is the diffusion mechanism. The resistance R_L represents the leakage current of the supercapacitor.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 645–652.

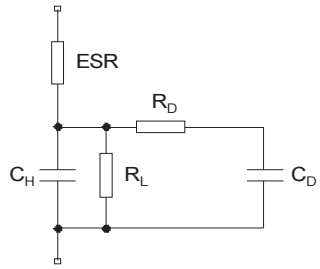

Fig. 3. A two-branch model of supercapacitor.

During stages 1 and 3, the charge is transported into and out of the structure, respectively. The charge is preserved in Helmholtz capacitance, as the value of R_D resistance is significantly higher than the value of ERS. In stage 2, when the charge stored in the structure is constant (the terminals are disconnected) we observe migration of the charge to pores that are less available and have not been yet occupied during the charging process. The time constant of this process is determined by diffusion resistance R_D and diffusion capacitance C_D (Fig. 3). Simplifying, the charge stored in Helmholtz capacitance C_H flows to diffusion capacitance C_D, which results in an overall voltage drop [10]. In this stage, also the leakage mechanism is responsible for the voltage drop. The leakage mechanism is dominant in this stage after relatively long time, while at the beginning it is the charge redistribution one that dominates [13].

In stage 4, when the supercapacitor is discharged and its terminals are open, an increase of voltage is observed between the terminals. It can be interpreted as an effect of a slow release of the charges stored deeply inside carbon pores. According to the electrical model (Fig. 3), after discharging Helmholtz capacitance C_H to zero volt and next opening the terminals, some amount of charge still remains in diffusion capacitance C_D because the resistance R_D is much higher than the ERS (the time constant $ESR \cdot C_H$ is smaller than the constant $R_D \cdot C_D$). Thus, the capacitance C_H will discharge faster than the capacitance C_D. Therefore, there will be a charge flow between C_H and C_D until the equilibrium state is reached [14].

When voltage is applied to the terminals of supercapacitor, ions migrate into vicinity of the electrode surface and form a Helmholtz plane. The electrode material is porous and ions migrate into pores, being forced by the electric field. Different size of pores and the random process of charging (penetration of pores at various speed) generates fluctuations in current flowing between the terminals when charged by a constant voltage supplied to the terminals. The fluctuations can be also observed during the discharging process when recording voltage fluctuations across the attached loading resistance.

The fluctuation phenomenon is induced by temperature (Johnson noise) but should exhibit some low frequency component ($1/f$-like noise) as well. That component should intensify when some areas of electrodes are on the verge of charging/discharging ability. We can assume that when some pores are blocked or almost blocked, the gathered charge can be removed (or stored) at a more slowly rate and low frequency fluctuations of that process should be observed. That phenomenon is observed in other electrochemical systems and applied to determine a corrosion rate [15].

Degradation of the supercapacitor as a result of operating conditions can be identified by an increase of ESR, a decrease of its electrical capacitance or by both mentioned changes. A decrease of capacitance is a result of blocking the pores by decomposed electrolyte and other chemical compounds. That degradation is irreversible. The pores can be also blocked by relatively large ions which exclude these pores from contributing to the supercapacitor terminal capacitance. These blocking processes are reversible and after some relaxation time can be at least partly restored .

Changes in the active area of carbon electrode (the number of active pores) should modify the intensity of fluctuation phenomenon. We can expect that the most intense 1/*f*-like noise is generated when the pores are on the verge of charging/discharging ability because such processes are rather very slow and will increase random components at a very low frequency range. When the pores are blocked completely, it means that these areas are excluded from any charging/discharging ability and noise generation as well. Thus, the 1/*f*-like noise should be potentially an interesting indicator of any process of pore blocking at its preliminary stage.

3. Low-frequency noise measurements in supercapacitor

A high value of capacitance C_H results in a very low frequency of low-pass filter formed by C_H, ESR and the loading resistance connected to the terminals of the tested supercapacitor. The electrical fluctuations are then filtered and only very low frequency components of noise generated inside the supercapacitor can be observed. This requires relatively long time of measurements because noise samples are recorded at very low rates. Moreover, estimation of power spectral density of noise samples requires averaging that lengthens the measurement process.

During the charging stage a stabilized current or voltage source is required to protect supercapacitor against overvoltage. These sources generate huge inherent noise or interference. It means that 1/*f*-like noise measurements during the charging stage could be overwhelmed by the inherent noise of the applied current or voltage source and therefore cannot be executed. Thus, we can assume that the low-frequency noise can be observed when the tested supercapacitor is fully charged (both capacitances C_H and C_D are completely charged) and some fluctuations are observed in the discharging current.

The supercapacitor can be discharged with a constant current or through a constant loading resistance. Discharging with a constant current shows the same limitations as charging with a constant current. An additional control unit is required which will introduce an additive noise source limiting identification of the 1/*f*-like noise component generated inside the discharged supercapacitor. That problem can be reduced when the supercapacitor is discharged through the joined constant loading resistance. A low-noise metallized resistor should be used for that aim. Moreover, when the supercapacitor is discharged through a resistance, the discharging time can be controlled by switching the loading resistor (*e.g.* between its low and high values). This method was applied for low-frequency noise measurements in the presented experimental studies.

A very low frequency noise component can be observed when the supercapacitor is discharged through a loading resistance securing a sufficiently low discharging current to record voltage fluctuations across the resistor for relatively long observation time. At the same time the discharging current should be huge enough to ensure intense voltage fluctuations across the loading resistor, up to the end of the recorded voltages.

A voltage across the loading resistor connected to the terminals of the charged capacitor is described by:

$$V(t) = V_0 \, \mathrm{e}^{(-t/RC)}, \tag{1}$$

where: V_0 is an initial voltage between the terminals of the charged supercapacitor; R is a loading resistor and C is a capacitance. The discharging time could be estimated by:

$$t = -RC \ln (V/V_0), \tag{2}$$

where: V is an acceptable voltage in the final stage of noise recording. For example, a supercapacitor of 2,5 F capacitance charged to a voltage of 2,7 V and next discharged by

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 645–652.

a loading resistor of 1 kΩ requires about 4 hours of discharging to reach a voltage between its terminals lower than 10 mV.

4. Measurement set-up

The measurement set-up for 1/*f*-like noise measurements in supercapacitor consists of (i) a current source with a voltage control, (ii) a switching unit with a set of keys (relays) to connect/disconnect supercapacitor to the elements of a bias circuit, (iii) a loading resistor and (iv) a data acquisition card (Fig. 4).

The controllable/programmable current source is used for charging/discharging of the examined supercapacitor before noise measurements. The current source is connected to the supercapacitor through the electronic keys in order to separate the supercapacitor and the current source during noise measurement. The keys are also used during measurements of charging/discharging currents and other parameters (*e.g.* a leakage current).

The data acquisition card provides dynamics and resolution of the recorded signals to measure AC (noise) and DC components of voltage ranging from a nominal voltage of the tested supercapacitor to nearly 0 V. The laboratory system secures 24-bit resolution and a voltage range of +/–10 V.

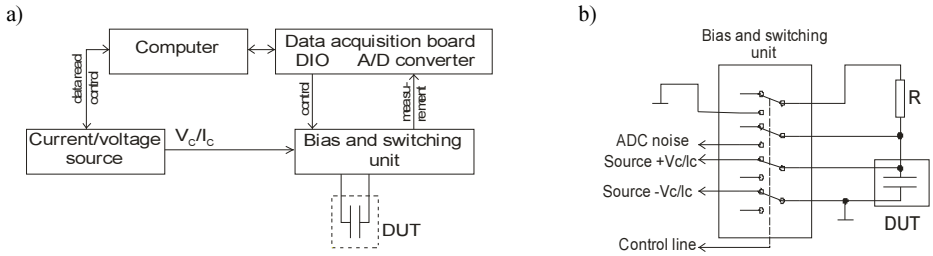

Fig. 4. The measurement setup: a block diagram (a); a switching and biasing unit (b).

Before noise measurements, the supercapacitor has to be charged to a given voltage. Moreover, its state should be stabilized. It is achieved either by leaving the supercapacitor with open terminals until its output voltage stabilizes (a voltage drop should be observed by a potentio-static cycle at a specific voltage until the current is sufficiently low, which indicates full charging of the tested supercapacitor – both C_H and C_D are fully charged). This operation is necessary to spread the charges within its structure. Next, the current/voltage source is disconnected and the loading resistor R and the data acquisition card are connected to the terminals of supercapacitor by the relay keys and voltage fluctuations across the loading resistor are recorded.

5. Experimental results and discussion

In the experiment, commercially available supercapacitors, DRL 2.7V 10F type, with a nominal capacitance 10 F and a nominal voltage 2.7 V, were used. A discharging curve of the tested supercapacitor discharged through the loading resistance 1 kΩ is shown in Fig. 5. An example of time record of voltage fluctuations and its histogram after removing the exponential trend are presented in Fig. 6.

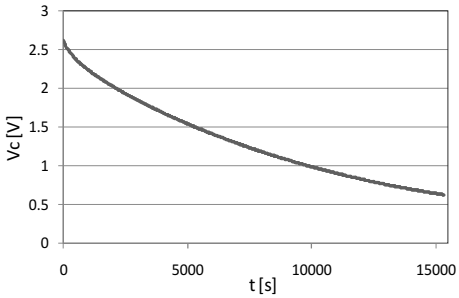

Fig. 5. A discharging curve of the tested supercapacitor, DRL 2.7V 10F type, through the loading resistance of 1 kΩ.

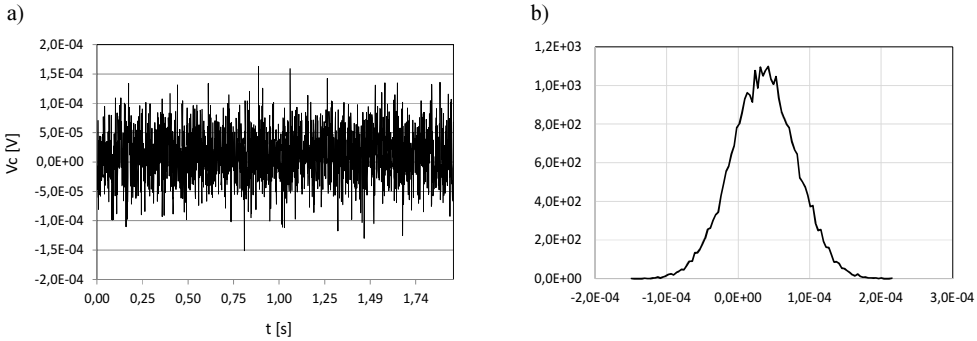

Fig. 6. An example of recorded time series of voltage fluctuations after removing the exponential trend (a) and its histogram (b).

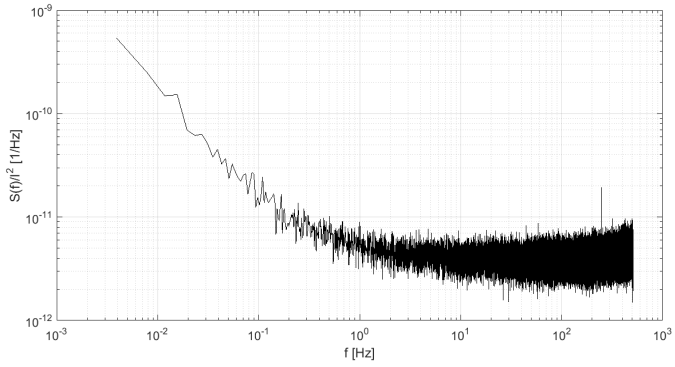

Fig. 7. Power spectral density of current fluctuations S(f) identified in the discharging current I of a DRL 2.7V 10F specimen when discharged through the loading resistance 1 kΩ. Power spectrum estimated in each time interval was normalized by the square of the discharging current.

The recorded samples were divided into sub-records and an FFT algorithm and the squaring operation was applied to estimate the power spectral density function S(f). The power spectra were normalized by the square of mean current I. The mean current value was calculated from the discharge current values in the time record the spectrum was calculated. Next, the spectra were averaged to reduce the estimation random error. An example of power spectrum density

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 645–652.

is shown in Fig. 7. It required about 4 hours of noise recording and the 1/*f*-like noise dominates at low frequencies below 1 Hz. The presented results confirmed that the 1/*f*-like noise can be observed in supercapacitors and used to assess their quality. The recorded noise exhibited the 1/*f* noise. We suppose that some degradation processes in the supercapacitor structure change its slope as in the case of 1/*f*-like noise generated in other porous materials for gas sensing [16]. Such information should be valuable for quality assessment of the tested supercapacitors.

It should be underlined that the presented preliminary results depend on quality of not only the tested supercapacitors but also the materials used for their preparation (carbon electrodes and the type of electrolyte). Therefore, we cannot assure that the proposed measurements will give satisfactory results of 1/*f*-like noise measurements in other types of supercapacitors.

6. Summary

Low-frequency noise generated in supercapacitors requires carefully selected experiment conditions and relatively long time of data recording – up to a few hours – to estimate its power spectral density. The 1/*f*-like noise prevails at frequencies below 1 Hz only in the examined commercial supercapacitors. The proposed and prepared laboratory measurement set-up controls the measurement time by changing the value of loading resistor. Its value is a compromise between either obtaining a too low discharging current and observing a very tiny noise component or obtaining a too big discharging current during very fast discharging, reducing time for noise recording and making necessary the operation of averaging to reduce the random error of power spectral density estimation. Additional long-term study is required to determine how 1/*f*-like noise is related to quality of the tested supercapacitor.

The next important issue is whether the 1/*f*-like noise can be observed in an almost discharged supercapacitor when a small current flows between capacitances C_H and C_D. It would be very interesting because it should shed light on processes occurring inside the supercapacitor's structure which has not been examined with the above presented method .

Acknowledgment

This research was financed by the National Science Center, Poland, project No. DEC-2014/15/B/ST4/04957, "Charging/discharging mechanism at the electrode/electrolyte interface of supercapacitors". Decision of 11.05.2017.

References

[1] Smulko, J., Darowicki, K., Zieliński, A. (2002). Detection of random transients caused by pitting corrosion. *Electrochimica acta*, 47(8), 1297–1303.

[2] Kiwilszo, M., Smulko, J. (2009). Pitting corrosion characterization by electrochemical noise measurements on asymmetric electrodes. *Journal of Solid State Electrochemistry*, 13(11), 1681–1686.

[3] Smulko, J. (2006). Methods of electrochemical noise analysis for investigation of corrosion processes. *Fluctuation and Noise Letters*, 6(2), R1–R9.

[4] Smulko, J., Kish, B., Granqvist, G. (2007). Quality assessments of electrochromic devices: the possible use of 1/*f* current noise. *Ionics*, 13(3), 179–182.

[5] Vandamme, L.K.J. (1994). Noise as a Diagnostic Tool for Quality and Reliability of Electronic Devices. *IEEE Transactions On Electron Devices*, 41(11), 2116–2187.

[6] Konczakowska, A. (1998). 1/*f* noise of electrolytic capacitors as a reliability indicator. *Quality and Reliability Engineering International*, 14, 83–85.

[7] Kötz, R., Carlen, M. (2000). Principles and applications of electrochemical capacitors. *Electrochimica Acta*, 45(15–16), 2483–2498.

[8] Beguin, F., Frąckowiak, E. (2013). *Supercapacitors:Materials, Systems and Applications*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

[9] Sedlakova, V., Sikula, J., Majzner, J., Sedlak, P., Kuparowitz, T., Buergler, B., Vasina, P. (2016). Supercapacitor Degradation Assesment by Power Cycling and Calendar Life Tests. *Metrol. Meas. Syst.*, 23(3), 345–358.

[10] Sedlakova, V., Sikula, J., Majzner, J., Sedlak, P., Kuparowitz, T., Buergler, B., Vasina, P. (2015). Supercapacitor equivalent electrical circuit model based on charges redistribution by diffusion. *Journal of Power Sources*, 286, 58–65.

[11] Szewczyk, A., Sikula, J., Sedlakova, V., Majzner, J., Sedlak, P., Kuparowitz, T. (2016). Voltage Dependence of Supercapacitor Capacitance. *Metrol. Meas. Syst.*, 23(3), 403–411.

[12] Yang, H., Zhang, Y. (2013). Analysis of Supercapacitor Energy Loss for Power Management in Environmentally Powered Wireless Sensor Nodes. *IEEE Transactions On Power Electronics*, 28(11), 5391–5403.

[13] Andreas, H.A. (2015). Self-Discharge in Electrochemical Capacitors: A Perspective Article. *Journal of The Electrochemical Society*, 162(5), A5047–A5053.

[14] Torregrossa, D., Bahramipanah, M., Namor, E., Cherkaoui, R., Paolone, M. (2014). Improvement of Dynamic Modeling of Supercapacitor by Residual Charge Effect Estimation. *IEEE Transactions On Industrial Electronics*, 61(3), 1345–1354.

[15] Smulko, J., Darowicki, K., Wysocki, P. (1998). Digital measurement system for electrochemical noise. *Polish Journal of Chemistry*, 72(7), 1237–1241.

[16] Lentka, Ł., Smulko, J.M., Ionescu, R., Granqvist, C.G., Kish, L.B. (2015). Determination of gas mixture components using fluctuation enhanced sensing and the LS-SVM regression algorithm. *Metrol. Meas. Syst.*, 22(3), 341–350.

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

ANALYSIS OF FREE-SPACE OPTICS DEVELOPMENT

Janusz Mikołajczyk[1)], Zbigniew Bielecki[1)], Maciej Bugajski[2)], Józef Piotrowski[3)], Jacek Wojtas[1)], Waldemar Gawron[3)], Dariusz Szabra[1)], Artur Prokopiuk[1)]

1) Military University of Technology, Institute of Optoelectronics, Gen. S. Kaliskiego 2, 00-908 Warsaw, Poland (janusz.mikolajczyk@wat.edu.pl, zbigniew.bielecki@wat.edu.pl, jacek.wojtas@wat.edu.pl, dariusz.szabra@wat.edu.pl, ✉ artur.prokopiuk@wat.edu.pl, +48 26 183 7740)

2) Institute of Electron Technology, Al. Lotnikow 32/46, 02-668 Warsaw, Poland (bugajski@ite.waw.pl)

3) VIGO System S.A., Poznańska 129/133, 05-850 Ożarow Mazowiecki, Poland (jpiotr@vigo.com.pl, wgawron@vigo.com.pl)

Abstract

The article presents state of work in technology of free-space optical communications (*Free Space Optics* – FSO). Both commercially available optical data links and their further development are described. The main elements and operation limiting factors of FSO systems have been identified. Additionally, analyses of FSO/RF hybrid systems application are included. The main aspects of *LasBITer* project related to such hybrid technology for security and defence applications are presented.

Keywords: optical communications, open path laser communications, line of sight communications, free space optics.

1. Introduction

During last decades FSO systems became an important direction of optoelectronic technology applications. FSO is also known as fibreless photonics. To obtain a broadband communication channel, *high frequency* (HF) modulated light pulses are used to transmit data through the atmosphere. These systems have been installed in terrestrial systems, as well as in systems for data transmission between space-space, space-earth and marine objects (Fig. 1). Operating in the infrared radiation spectrum, the FSO can provide links with a very high data rate (tens of Gigabits per second) between various platforms offering ranges of several kilometres near the sea level or even over 100 km at high altitude.

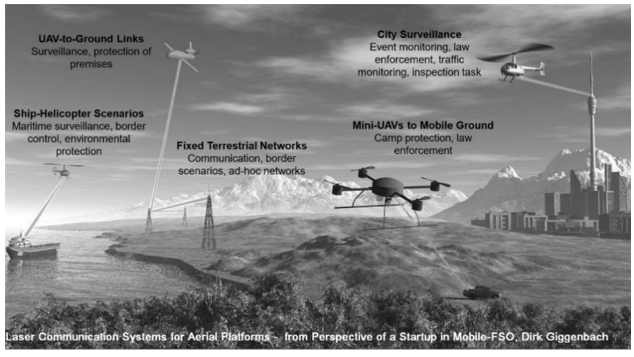

Fig. 1. Examples of scenarios of the FSO systems application.

Article history: received on Jun. 14, 2017; accepted on Sep. 02, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0060.

Important FSO systems' applications are *e.g.*: ad-hoc mobile communications during crisis situations, so-called "last mile" links in urban areas, a special data transfer *e.g.* between ships and land, and secret military data links. It was shown that a command centre would be able to send information to vehicles or soldiers at a distance of "line-of-sight" with a transmission rate from 25 Mbps to 1.25 Gbps [1].

The most important advantages of FSO technology are:
- use of radiation spectra is not covered by formal regulation;
- high transmission rates up to 10 Gbps [2];
- no interferences with other transmissions (insensitivity to EM interference);
- fast, low-cost and easy installation;
- high immunity to interception and jamming;
- user-adjusted capacity with an option of data link reconfiguration;
- commercial availability.

The main disadvantage of the FSO system is its high sensitivity on weather conditions (atmospheric attenuations). Different weather phenomena like fog, snow and rain, turbulence, are able to scatter and to absorb the optical signal. As a result, both range and data rate of data transmission channel are reduced. To minimize the influence of these negative factors, characterization of various weather conditions and selection of so-called atmosphere transmission windows are required [3]. The recent developments in optoelectronic technology make it possible to construct some FSO systems alternative to RF wireless ones in mainstream communication applications.

The paper is arranged as follows. In Section 2, the fundamental knowledge of FSO system construction is given. The influence of selected atmospheric effects and other factors on FSO link performance is analysed in Sections 3 and 4. A description related to the selection of a radiation wavelength for FSO communication is included in Section 5. The research results of the FSO systems designed at the Institute of Optoelectronics, MUT are presented in Sections 6, 7 and 8, together with a concept of FSO/RF hybrid system based on the "LaserBITer" project, financed by the Polish National Centre of Research and Development .

2. Free Space Optics

In general, an FSO link consists of an optical signal transmitter and a receiver. As shown in Fig. 2, the transmitter is used to transmit data signals in free space by modulation of optical radiation. Its main elements are a radiation source, a laser modulator and optical devices.

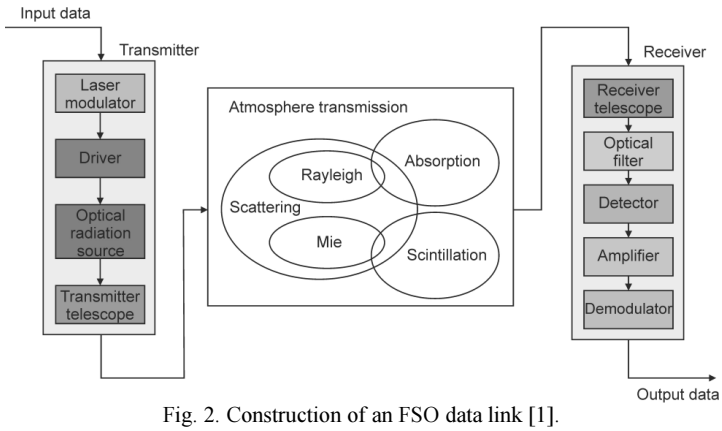

Fig. 2. Construction of an FSO data link [1].

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

The laser modulator modulates the optical signal with an electrical one by varying *e.g.* the laser biasing current. The most popular optical device is a telescope. It is applied to direct optical radiation towards the receiver.

The optical signal is attenuated in the atmosphere by absorption, scattering, scintillation, propagation geometrics and other phenomena. In practice, the total radiation attenuation $A(\lambda)$ caused by atmosphere can be calculated as:

$$A(\lambda) = \alpha_{\text{fog}}(\lambda) + \alpha_{\text{snow}}(\lambda) + \alpha_{\text{rain}}(\lambda) + \alpha_{\text{scattering}}(\lambda), \text{ [dB/km]}, \quad (1)$$

where $\alpha_x(\lambda)$ is attenuation caused by the mentioned weather conditions and λ is an operational wavelength.

The receiver usually consists of a telescope, an optical filter, a photodetector, a preamplifier (preamp), and a demodulator for proper retrieving the information signal. The telescope collects and focuses the optical radiation on the photodetector active area. The filter reduces the background radiation (*e.g.* solar illumination). The photodetector converts the photon energy into an electrical signal. It should provide a high responsivity at the wavelength of interest, a small value of noise, sufficient values of dynamic range and signal bandwidth. The most commonly used photodetectors are a pin photodiode and an *avalanche photodiode* (APD). The photodetector output signal is amplified by a special construction of preamp. During the last procedure, the amplified signal is analysed using the demodulator.

To determine performance of FSO link, many factors should be taken into account, *e.g.* the operation wavelength, the light source power, the beam divergence angle, the photodetector detectivity and the aperture diameters of applied optical devices.

The beam size at the receiver surface is dependent on the beam divergence and the transmission range. Typically, the beam divergence is used in a range from 1 mrad to 8 mrad, although in some special links the values from 6 µrad up to 180° are also applied [5, 6].

The selection of a light source for FSO applications depends on various factors. The most important are: the radiation pulse power, modulation capabilities, lifetime, eye safety, beam size and divergence angle, physical dimensions, compatibility with other transmission media, price and purchase availability. The parameter values of some radiation sources applied in selected FSO systems are listed in Table 1.

Table 1. Light sources applied in selected FSO systems.

Laser	Wavelength [nm]	Laser/LED power	Beam divergence	Application	Data source
Matrics LEDs	450	6 W	180°	Underwater communication	www.sonardyne.com
Nd:YAG	532	250 mJ 12 ns	110 µrad	Deep space mission	[7]
LD	532/486	5 W	180°	Underwater communication	www.saphotonics.com
LD	785	25 mW	1 mrad	Ethernet	www.geodesy-fso.com
AlGaAs	830	60 mW	6 µrad	Inter-satellite communication	[5]
Argon-ion/GaAs	830	13 W	20 µrad	Ground-to-satellite link	[8]
VCSEL	850	9 mW	3.5 mrad	Last mile link	www.polixel.pl
LED	800–900	bd	17 mrad	Communication between buildings	freespaceoptics.ca
LD	1550	113 mW	50 mrad	UAV-to-UAV link, L = 2km	[9]
LD	1550	200 mW	19.5 µrad	Ground-to-UAV link	[10]
QCL	8400	740 mW (100ns, f = 1 MHz)	2 mrad	Laboratory FSO link (IOE MUT)	[11]

Laser diodes (LDs) are typically used in current commercially available FSO systems. However, in some FSO constructions non-lasing sources such as *light-emitting diodes* (LEDs) or *IR-emitting diodes* (IREDs) are also applied. But – compared with LEDs – LDs are characterized by a higher output power, energy efficiency, modulation rate, and by a less diverged beam. Laser diodes called *vertical emitting lasers* (VCSELs) are high-speed radiation sources that ideally suit for high-speed (Gbps) data communication. These lasers are also characterized by very low threshold currents, non-stringent requirements for the modulation signal, and a good beam quality. They are relatively stable and therefore do not require power control using a photodiode monitor. The most common VCSELs are composed of GaAs/AlGaAs to emit light in a range of 750–980 nm.

Recently, *quantum cascade lasers* (QCLs) are used as infrared radiation sources ($\lambda \sim 3.5 \div 24$ μm) basing on the unipolar lasing mechanism; QCLs have unique high-frequency characteristics with theoretical bandwidths above 100 GHz [12].

In addition, there has been also developed a high-bandwidth underwater communication system using blue-green lasers. For example, the BlueComm modem family constructed by Sonardyne company provides data transmission rates exceeding 500 Mbps. A transmission rate of 1 Gbps at distances greater than 150 m was also reported [13].

In practice, there are four common topologies of FSO networks: point-to-point, point-to-multipoint, mesh, and ring ones. However, these topologies can be combined. In the point-to-point arrangement, a transmission rate from 155 Mbps to 10 Gbps at a distance from 2 km to 4 km can be obtained. Such a link provides a dedicated connection with a higher bandwidth, but it is not cost-effectively. In comparison, the point-to-multipoint configuration is cheaper but offers a worse bandwidth (the same data rate at a distance from 1 km to 2 km). The mesh topology is able to transmit data with a rate of 622 Mbps at shorter distances from 200 m to 450 m. The ring topology is usually used in metropolitan networks. There are "backbones" represented by fibre or FSO high-speed rings [14].

A roadmap of FSO technology is shown in Fig. 3. It can be noticed that since 1990s there have been commercially available FSO systems with data rates of up to several tens of Mbps. Nowadays, this rate has been gradually increasing up to 10 Gbps.

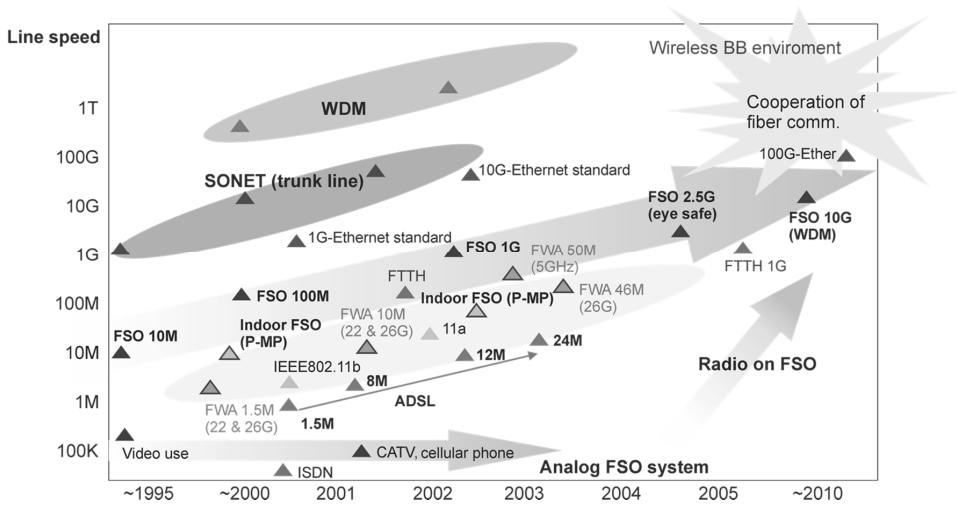

Fig. 3. A roadmap of FSO technology [15].

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

3. Atmospheric effects

The atmosphere interactions with optical radiation due some phenomena depend on its composition. Practically, the atmosphere consists of different molecular species and small particles like aerosols (fog, forest exudates, dust, sea-salt particles, soil particles, volcano debris, particulate, air pollutants, smog and smoke), ice particles, and water droplets. Thus, the atmosphere causes attenuation of optical signals by absorption, scattering, and scintillation. Nonetheless, atmospheric transmission windows are defined by the molecular absorption, that is a spectral selective phenomenon (Fig. 4) [16]. In general, an atmospheric attenuation τ is described by Beer's law:

$$\tau = exp\left[-\left(\alpha_{abs} + \beta_{scat}\right)L\right], \tag{2}$$

where: L is a distance between the transmitter and receiver; α_{abs} and β_{scat} are coefficients of atmosphere absorption and scattering, respectively.

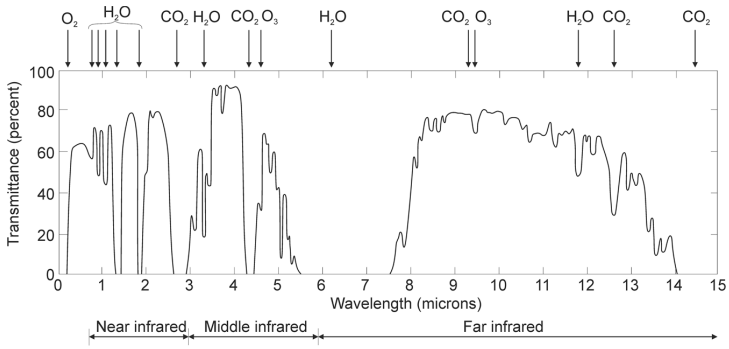

Fig. 4. Transmittance through the atmosphere as a function of wavelength.

Absorption is caused by atmospheric molecules, the energy levels of which can be excited by incident photons. An absorption coefficient depends on the type, effective absorption cross-section σ_{abs} and concentration N_{abs} of gas molecules. These parameters are related as follows [17]:

$$\alpha_{abs} = \sigma_{abs} N_{abs}. \tag{3}$$

The scattering process causes propagation of the redirected radiation beam propagates in various directions different from the original one [2]. There are three main types of scattering: Rayleigh, Mie, and non-selective. The scattering type depends on the relationship between the size of scattering particles and the wavelength of propagated light. *Rayleigh scattering* is caused by particles with a size much smaller than the light wavelength. In this case, the scattering intensity decreases with the wavelength as $\sim\lambda^{-4}$. When the particle size is comparable with or is as large as the radiation wavelength, *Mie scattering* is observed. *Non-selective scattering* occurs for particles' sizes greater than the beam wavelength. In this case, the Mie theory is approximated by the principles of reflection, refraction and diffraction. The scattering coefficient depends on the concentration N_{scat} and effective cross-section σ_{scat} parameter values of the particles and can be described by:

$$\beta_{scat} = \sigma_{scat} N_{scat}. \tag{4}$$

The total scattering coefficient is given by:

$$\beta_{scat} = \beta_m + \beta_a, \tag{5}$$

where β_m and β_a denote Rayleigh (molecular) and Mie (aerosols) scattering.

The Rayleigh scattering coefficient is given by:

$$\beta_m = \sigma_m N_m, \tag{6}$$

where: σ_m is a Rayleigh scattering cross-section; N_m is a volumetric density of air molecules.

The parameter σ_m depends on the index of refraction n, volumetric density of the molecules N, and the radiation wavelength:

$$\sigma_m = \frac{8\pi^3\left(n^2-1\right)^2}{3N^2\lambda^4}. \tag{7}$$

Rayleigh scattering is significant in the ultraviolet and visible spectral ranges. Moreover, it is negligible in the infrared range. The Mie scattering coefficient is expressed by:

$$\beta_a = \sigma_a N_a, \tag{8}$$

where: σ_a is a Mie scattering cross-section; and N_a is a volumetric density of air particles.

The value of coefficient β_a can be estimated by visibility V with the expression[1]:

$$\beta_a = \left(\frac{3.91}{V}\right)\left(\frac{0.55}{\lambda}\right)^{\delta}, \tag{9}$$

where: δ is a coefficient with a value between 0.7 and 1.6 corresponding to visibility conditions given in km; λ is a wavelength of propagating beam (μm) [19].

The fog particles float in the air for a longer time than rain droplets. Additionally, they are characterized by a size smaller than the radiation wavelength. Thus, the scattering due to rainfall (non-selective scattering) is less effective than to fog (Mie scattering). The rain scattering coefficient can be determined using the Stroke Law [20]:

$$\beta_{rain\ scat} = \pi r^2 N_a Q_{scat}\left(\frac{r}{\lambda}\right), \tag{10}$$

where: r is a radius of raindrop (from 0.001 cm to 0.1 cm); N_a is a distribution of rain drops, and Q_{scat} is a scattering efficiency.

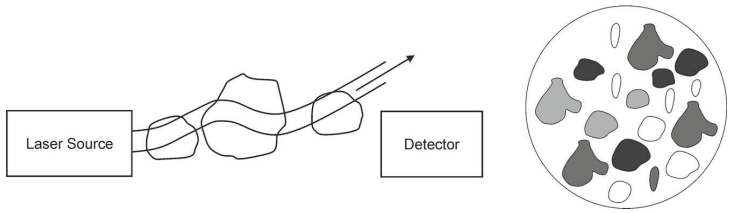

Fig. 5. Optical path changes due to turbulence: eddies are larger than the beam diameter, with scintillation spots and aperture averaged.

Another important atmospheric factor that limits FSO data link capabilities is turbulence. *Clear-air turbulence* (CAT) is defined as chaotic streams and eddies of air masses during the absence of any clouds. Such air movements are described as turbulent ones, because air masses are moving at widely different speeds [21]. As a result of this phenomenon phase shifts of the propagated optical radiation are noticed. The distortions of front-wave can be observed as intensity changes referred as scintillation. Aerosols, moisture, temperature and pressure

[1] According to the Kruse model, visibility is defined as a path length, where radiation of 550 nm is attenuated to 0.02 of its original value and it is estimated through observation.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

fluctuations produce variations of the air density and thus also its refractive index [3]. Air eddies can bend the optical path if their size is larger than the beam diameter (Fig. 5). In the opposite situation, constructive and destructive interferences are created, resulting in temporal fluctuations of light intensity (spots) at the receiver surface.

For scintillation scaling, a *refractive index structure parameter* C_n^2 is introduced into calculations. A number of parametric models have been formulated to describe the C_n^2 profile. One of the most commonly used model is described by Hufnagel-Valley [23]:

$$C_n^2(h) = 0.00594\left(\frac{v}{27}\right)^2 (10^{-5}h)^{10} exp\left(-\frac{h}{1000}\right) + 2.7\cdot 10^{-16} exp\left(-\frac{h}{1500}\right) + A_0 exp\left(-\frac{h}{100}\right), \tag{11}$$

where: h is an altitude in m; v is a wind speed at high altitude in m/s; A_0 is a turbulence strength on the ground level; $A_0 = 1.7\cdot 10^{-14}\text{m}^{-2/3}$. In practice, this parameter depends also on the geographical location and time of day. There are three different turbulence effects: *scintillation, beam wander* and *beam spreading*. *Scintillation* is the most important for FSO links, causing intensity fluctuations at the receiver surface. The level of scintillation can be measured in terms of the irradiance variance given by:

$$\sigma_i^2 = 1.23 C_n^2 k^{\frac{7}{6}} L^{\frac{11}{6}}, \tag{12}$$

where $k = 2\pi/\lambda$ is the wave number.

The variance is linearly proportional to C_n^2, nearly proportional to both $1/\lambda$ and the square of the link distance [24]. Therefore, systems of shorter wavelengths have a proportionally higher variance caused by scintillations. This effect increases with the data range and becomes more critical for small aperture photo-receivers [25].

Beam wandering, as well as the scintillation index, is an important characteristic of radiation propagation. It determines requirements for tracking and pointing instruments of FSO system [26]. This effect is observed as a random movement of the focused beam on the photodetector surface. The beam wandering is also expressed in terms of local fluctuations of the irradiance intensity. It results in an increase of system *bit error rate* (BER) and, appropriately, the tracking error. Recently, many studies indicated that partly coherent beams are less affected by the turbulence than the fully coherent beams. So, the use of a partly coherent beam source reduces the radiation intensity fluctuation at the receiver [27].

Beam spreading is related to the broadening of the beam size at the receiver surface beyond an expected pure diffraction limit. Fig. 6 shows the laser beam propagation through the turbulent atmosphere.

The additional laser beam spreading caused by turbulence grows with the increase of both refractive index structure parameter C_n and propagation length (Fig. 7). This spreading is expressed as:

$$T_{beam} = 1.33\sigma_I^2 \Gamma^{5/6}, \tag{13}$$

while Γ is given by:

$$\Gamma = \frac{2\lambda L}{2\pi\omega^2(l)}, \tag{14}$$

where: L is a distance from the source; $\omega^2(l)$ is an initial beam waist at $L = 0$. The parameter σ_I defined as a beam amplitude or irradiance equals:

$$\sigma_I^2 = 1.23 C_n^2 \left(\frac{2\pi}{\lambda}\right)^{7/6} L^{11/6}. \tag{15}$$

In the latest FSO systems, some techniques are applied to mitigate such atmospheric effects as scintillation or beam wander. These techniques use *e.g. adaptive optics* (AO), diversity techniques, aperture averaging, and fast tracking antennas.

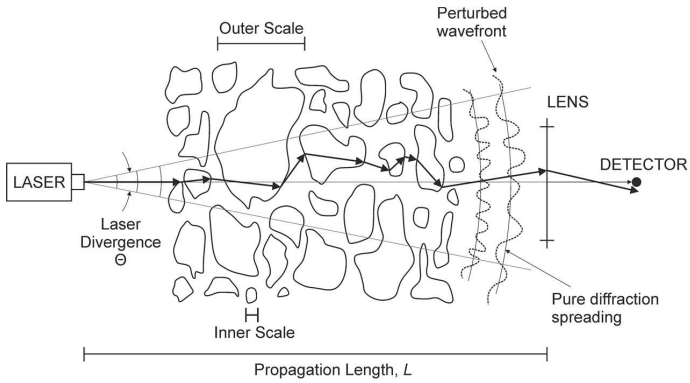

Fig. 6. Propagation of a laser beam through the turbulent atmosphere [28].

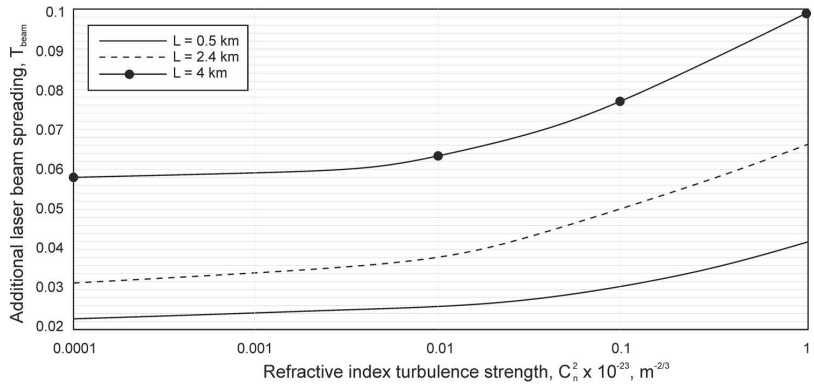

Fig. 7. Laser beam spreading vs. refractive index structure parameter for different propagation lengths and a beam wavelength of 1550 nm [28].

Adaptive optics are designed to continuously measure and correct wave-front errors. Beam diversity can occur in some forms:

- spatial – requiring multiple transmitters and receivers;
- temporal – requiring double transmitted signals, separated by a time delay;
- wavelength – requiring at least two different wavelengths transmission of data.

Numerous methods for fine tracking and automatic acquisition have been developed. These methods include the use of quadrant-detectors, servo-motors, voice-coils, stepping-motors, mirrors, CCD arrays and MEMS [29].

Practical performance of FSO systems is also limited by *geometric losses*. These losses occur when the light beam spreads to a size larger than the receiver's aperture. Geometric loss is expressed by a ratio of the receiver's aperture diameter D_R and irradiated beam diameter D_T. This ratio can be determined by the formula:

$$geometric\ loss = \frac{(D_R)^2}{(D_T + L\theta)^2}, \tag{16}$$

where θ – a beam divergence [mrad]. In general, a system is perfectly aligned when the centre of the Gaussian power distribution is at the optical axis of the receiver.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

4. BER, SNR and optical link budget

Taking into account atmospheric attenuation and geometric loss, a radiation power P_R registered by the receiver is given by [30]:

$$P_R = P_T \frac{D_R^2}{(D_R + L\theta)^2} exp(-\tau L), \tag{17}$$

where: P_T is a power emitted by the transmitter; τ is a total coefficient of atmospheric attenuation.

In digital transmission, the BER value is determined as a ratio of the number of bit errors and the total number of transmitted bits during a studied time interval [31]. For FSO communication, it can be described as [32]:

$$BER = \frac{1}{2} erfc\left(\frac{RP_R}{2\sqrt{2N^2}}\right), \tag{18}$$

where: *erfc* is called a Gauss error function; R is a detector responsivity; and N is thermal noise of the receiver. The value of BER also depends on the modulation scheme. For FSO links with the on-off keying modulation, the BER performance is characterized by a *signal to noise ratio* (SNR) [33]:

$$BER = \frac{exp(-SNR/2)}{(2\pi SNR)^{0.5}}, \tag{19}$$

Taking into account turbulence, the SNR is estimated [22]:

$$SNR = \left(0.31 C_n^2 k^{7/6} I^{11/6}\right)^{-1}, \tag{20}$$

where $I = |E^2|$ is radiation irradiance.

Summarizing, the received optical power can be calculated as [34]:

$$P_R = P_T G_T \eta_T L_{PT} L_{FS} G_R \eta_R L_{PR} L_A, \tag{21}$$

where: η_T, η_R are losses that include imperfect optical components of both transmitter and receiver; $G_T = (\pi D_T/\lambda)^2$ is a gain of the transmitting aperture; $L_{PT} = exp\left(-8\theta_{jit}^2/\theta^2\right)$ is a pointing loss of the transmitter; $L_{FS} = (\lambda/4\pi L)^2$ is a free-space propagation loss; $G_R = (\pi D_R/\lambda)^2$ is a gain of the receiving aperture; L_{PR} is a pointing loss of the receiver; L_A is atmospheric attenuation at the operating wavelength; θ_{jit} is an optical beam jitter angle; and θ is an optical beam divergence as set by diffraction.

It should be also noticed that the application of a high-sensitive photo-receiver in an FSO system with a large-aperture lens makes it possible to increase the influence of the background radiation on the data signal. Sometimes, direct sunlight may cause link outages for a period of time. In these circumstances, narrowing the photo-receiver FOV and using a narrow-bandwidth optical filter can improve the system performance.

5. Wavelength selection for FSO communications

The selection of a wavelength for the FSO data link is a very important issue. Nowadays, commercial FSO systems usually operate in the spectra of 780–850 nm and 1520–1600 nm. To determine a wavelength range, it should be taken into consideration the availability of the main FSO components defined by their transmission range, eye safety, modulation rate, costs, and so on. The eye safety is one of the most important restrictions to the optical power level emitted by an FSO transmitter. Lasers emitting radiation at a wavelength of 1550 nm or about 10000 nm can be used more safely than those with 850 nm and 780 nm. This is due to the fact that infrared

radiation with wavelengths above 1400 nm is absorbed by the transparent parts of the human eye before reaching the retina. That is why the *maximum permissible exposure* (MPE) for these wavelengths is higher than for shorter ones.

The International *Electro-technical Commission* (IEC) classifies lasers into four safety classes depending upon their beam power, wavelength and possible hazards [35]. Most of the FSO systems use Class 1 and Class 1M lasers. For example, an FSO system operating at a wavelength of 1500 nm can transmit a light beam with 50 times higher power comparing with a system working at a shorter wavelength range. It enables to propagate radiation over longer distances in the case of worse weather, and to support higher data rates [30].

In the near-infrared spectrum (NIR, 780÷850 nm), reliable, inexpensive, high-performance optoelectronic devices, *i.e.* lasers and detectors, are readily available and commonly used in the FSO transmission equipment. Advanced VCSEL lasers and silicon photodiodes are used for operation at this wavelength. Si detectors typically have the maximum of their spectral responsivity near the value of 850 nm, making in conjunction with VCSELs very efficient tools. Silicon detectors are ideal for FSO systems operating at a very high bandwidth – 10 Gbps.

The *short-wavelength-infrared spectrum* (SWIR, 1520–1600 nm) is also well applicable for FSO links. High quality lasers and detectors are readily available. These wavelengths are also used in the fibre technology. As radiation sources, Fabry-Perot and *Distributed-Feedback lasers* (DFB) based on InGaAs/InP semiconductor technology are used. For construction of an FSO receiver, InGaAs detectors are usually applied. These detectors based on PIN or APD technology are optimized for operation at the wavelength of 1310 nm or 1550 nm providing a data rate of 10 Gbps.

The *long-wavelength-infrared spectrum* (LWIR) FSO systems are more challenging because of practical aspects. However, in this spectral range, there is observed a smaller impact of absorption and scattering on beam propagation through moderate fog comparing with that of other infrared ranges. Also, atmospheric turbulence is characterized by a smaller impact on transmission. Additionally, a smaller influence of solar radiation (29 dB) at a wavelength of 10 μm, comparing with the wavelength of 1550 nm is noticed [36, 37]. Recently developed *quantum cascade lasers* (QCL) are very attractive radiation sources operating in this spectral range. They are compact, high-power semiconductor lasers with the frequency characteristics of even 100 GHz bandwidth. This makes QCLs important tools for constructing communication systems. In the case of FSO receiver design, an MCT detector characterized by ultra-high detectivity and GHz signal bandwidth is applied.

Summarizing, there are three different optical radiation spectra employed in FSO systems. In their construction, different radiation sources and detectors are used. These optoelectronic elements are characterized by parameters dependent on wavelengths and data rates. That is why the analyses of FSO system design should take into consideration the impact of the operation spectra on the atmosphere transmission. For example, Fig. 8a shows the total attenuation versus low visibility at wavelengths of 780 nm, 850 nm and 1550 nm. These wavelengths correspond to the operation spectral ranges of commercially available FSO systems. The total attenuation at a wavelength of 1550 nm is lower than at others. Therefore, to reduce the beam attenuation during hazy days, the SWIR-FSO system should be used. It is observed that the radiation attenuation increases with the link range (Fig. 8b). For a radiation wavelength of 1550 nm, the growth speed is lower than for others.

The performed analyses show that the atmospheric attenuation depends also on the rainfall rate (Fig. 9). However, rain has not so strong spectral influence, because raindrops have larger size compared with laser wavelengths causing minimal scattering of light beam.

Nowadays, the dynamic development of LWIR-FSO systems is observed. But these works are mainly performed at lab-experiment level. In Fig. 10 a comparison of LWIR radiation

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

attenuation with that obtained for other ranges for different values of visibility is presented. It is shown that the LWIR range is characterized by better transmission in a wide range of visibility.

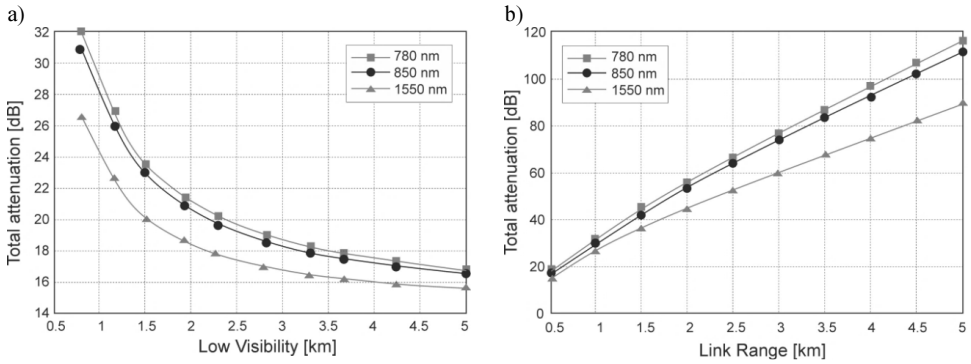

Fig. 8. Total attenuation versus average visibility (a); and total attenuation versus link range for different wavelengths (b).

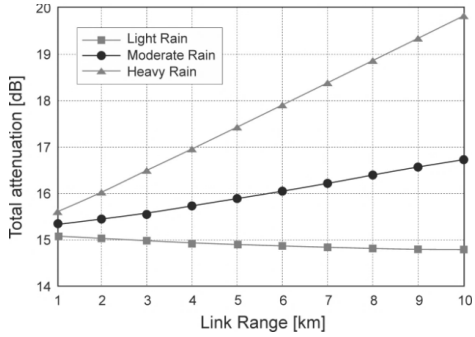

Fig. 9. Total attenuation versus link range for different rainfall rates.

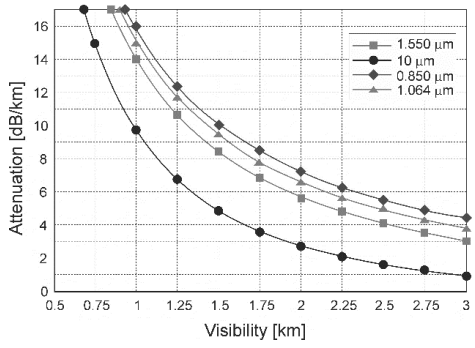

Fig. 10. Radiation attenuation versus visibility [38].

There are only a few reports describing results of experimental verification of beam transmission at the wavelengths of interest. Fig. 11a shows transmission losses for four different wavelength links as a function of time.

There was observed a significant decrease in transmission for three shorter wavelengths during the day time. In the same time, an increase of water vapour concentration was also registered. Similar experiments were performed for different oil vapour concentrations. In this situation, the LWIR radiation is also characterized by the lowest attenuation (Fig. 11b).

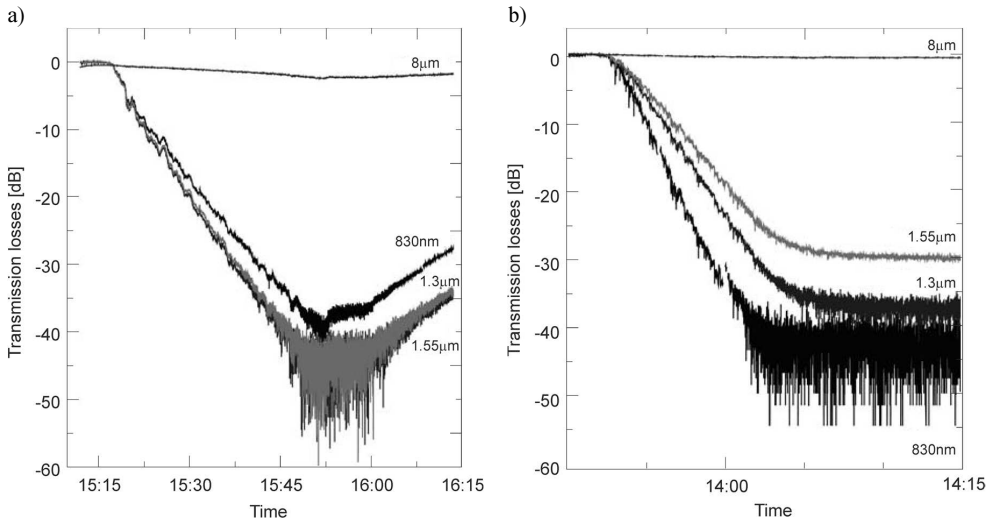

Fig. 11. Transmission losses for four different wavelengths of FSO links: as a function of time during which the water vapour concentration increased (a); as a function of increased oil vapour concentration (b) [39].

6. FSO/RF hybrid data link

Free Space Optics is sensitive to atmospheric conditions reducing visibility, *e.g.* precipitation, fog, haze, or scintillation. These factors may limit the data link range to a few km or several hundred metres. It is also known that RF links show good transmission in fog, but high attenuation in the presence of precipitations like rain and snow. This results in a better matching the RF wavelength to sizes of *e.g.* rain droplets. However, RF channels are also more susceptible to interference and jamming.

Combining these two technologies into one FSO/RF hybrid link may increase the data transfer availability and ensure higher security keeping a high speed of transmission. Such a hybrid construction can be classified into three categories: redundant systems, switch-over systems and load-balancing systems [40]. The redundant systems duplicate data and transmit them simultaneously over both the FSO and RF data links. In contrast, the switch-over systems transmit data using only one link. Usually, the FSO link is chosen as the primary link whereas the RF one operates as a backup. In practice, the RF link compensates the reduced bandwidth of FSO link during bad-weather conditions. The load-balancing systems distribute the data traffic between the FSO and RF links according to the connectivity quality, thus exploiting the full available bandwidth at each time.

The hybrid FSO/RF technology is especially dedicated to military communication systems, crisis management, intelligent transportation systems, and telemetry. In the military area, it can be used in transmission systems at Tactical Operations Centres, airborne networks, cross-links between satellites, as well as in different types of platforms: space-to-air, space-to ground and air-to-ground [41].

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

7. Free Space Optics technology at the Institute of Optoelectronics, MUT

In the Institute of Optoelectronics, the research related to laser communication systems started in 1990s. The first FSO system operating at a wavelength of 850 nm was developed in 1993 and provided a data transfer rate of 10 kbps. In 2004, there was constructed the second system with a transfer rate of 100 Mbps and an operation wavelength of 1.54 µm. The work on LWIR-FSO links was started in 2006. That link used QCL's laser system from Cascade Technologies. The receiver consisted of an off-axis mirror system from Janes Technology and a low-noise MCT detection module from VIGO System S.A. The main limiting factors of its data transmission rate were the modulation bandwidth and the duty cycle of generated pulses.

Thanks to the recent progress in semiconductor technology, QC lasers with a much higher power, repetition rate and duty cycle of pulses have been constructed. In 2009, it was decided to design the second model of LWIR FSO link. In that construction, a laser system of Alpes Lasers SA, a germane lens and optimized MCT detection modules from VIGO System S.A were used. To connect the optical link with the data network, a fully programmable RCM 4200 module operated by a Rabbit 4000 processor with a complete Ethernet interface was applied. The module worked as a buffer for receiving data frames from Ethernet network, performing data analysing and validation. Parameters of the two LWIR-FSO systems are listed in Table 2.

Table 2. Parameters of the constructed LWIR-FSO data links.

Parameter	First model (2006)	Second model (2009)
Operation wavelength	10 µm	8.4 µm
Pulse peak power	100 mW	200 mW
Detection module detectivity	$3.2 \cdot 10^9$ cmHz$^{1/2}$/W	$3 \cdot 10^{10}$ cmHz$^{1/2}$/W
Beam divergence	1.5mrad	2.5 mrad
Data rate	115 kbps	2 Mbps
Range	1.5 km (Vis = 2 km)	2.5 km (Vis = 2 km)

8. LasBITer Project of radio-optical data link

LasBITer is a hybrid data link design consisting of optical and radio communication channels (Fig. 12). The FSO transmitter project consists of a compact laser head with a QC laser constructed at the Institute of Electron Technology, a laser driving unit, a temperature controller and a parabolic off-axis mirror optical device. Additionally, the device is equipped with a laser power monitoring system.

8.1. Project description

The FSO receiver is built of an optimized optical system and a detection module from VIGO System S.A. The optical system provides both high data rate and good conditions to start the FSO data link and to minimize the influence of turbulence on the link availability. The detection module is equipped with an MCT detector and a pulse power control unit.

The communication unit enables data transmission using the designed FSO system and a commercial RF. It also enables measuring a bit error rate in each of these channels. It is based on XILINX FPGAs technology. This application enables flexible changing of both data coding and configuration of forward error correction methods. In this unit, a control data stream is added to the useful information in order to monitor the link quality.

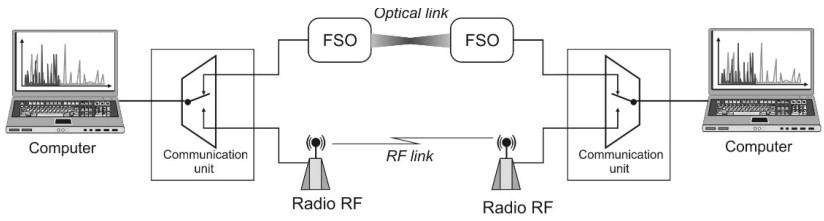

Fig. 12. A concept of the FSO/RF hybrid data link.

A special controller of FSO data link is also developed to control the parameters of its key-components (power supplies, radiation power, operation temperature, *etc*.). The radio data link is based on commercial transmission modules operating at the military frequency of 1.4 GHz. Two modes of co-operation of FSO-LWIR and RF are designed. In the independent operation mode, the data link user decides which of the channels will be used for data transmission. It is also possible to distribute the data proportionally to the configured data rate, so-called load balancing, using the second mode of its operation.

Compared with some previously reported hybrid systems [42, 43], the *LasBITer* data link design has a better data range, communication availability, and transfer data safety.

The goals should be obtained by the use of dedicated 8÷12 μm quantum cascade lasers and optimized MCT detection modules.

8.2. Quantum cascade lasers

The quantum cascade lasers are unipolar devices based on tunnelling and inter-sub-band transitions, in which the electronic states, wave functions and lifetimes of relevant states are engineered through the quantum mechanical confinement imposed by a complex multilayer structure. The second main feature of this type of lasers is the cascading scheme of carriers' route through the laser active region. That means that a single carrier is used more than one time for generating a photon carrier. For QCLs operation, an extremely precise tailoring of energy levels of quantum states, optical dipole matrix elements, tunnelling times and scattering rates of carriers is required. The physical basis of QCLs operation is fundamentally different from that of classical bipolar semiconductor lasers, in which the emission is due to the inter-band radiative recombination of pairs of carriers instead of inter-sub-band transitions which lead to lasing in QCLs [44].

At this moment, a wavelength range of QCL radiation spans from ~3.5 μm up to ~250 μm. So it generally covers a very wide infrared spectrum, from mid-IR up to far-IR. In comparison with the performance of bipolar lasers, this one provides about two orders of magnitude increase of the wavelength range available for semiconductor lasers, towards the longer wavelengths. The huge spectral flexibility of the emission is a result of the application of the intra-band generation mechanism.

For the purpose of *LasBITer* data link project, the lattice matched (~9.2÷9.4 μm) $Al_{0.48}In_{0.52}As/In_{0.53}Ga_{0.47}As/InP$ QCLs technology has been selected [45, 46]. The laser structures consisted of 30- segments. The active region of the lasers was of a 4-well 2-phonon resonance design. The layer sequence of one period of the structure, in nanometres, starting from the injection barrier is: **4.0**, 1.9, **0.7**, 5.8, **0.9**, 5.7, **0.9**, 5.0, **2.2**, 3.4, **1.4**, 3.3, **1.3**, 3.2, **1.5**, 3.1, **1.9,** 3.0, **2.3**, 2.9, **2.5, 2.9** nm. The AlInAs layers are printed in bold. The total thickness of one period is 59.8 nm. The underlined layers are *n* doped to $2.0 \times 10^{11}cm^{-2}$. The conduction band profile and squared wave-functions of moduli in the injector/active/injector segment of the $Al_{0.48}In_{0.52}As/In_{0.53}Ga_{0.47}As/InP$ laser under the applied field of 50 kV/cm are shown in Fig. 13.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

The electronic band structure of QCL has been calculated by solving the Schrödinger equation with position-dependent effective mass.

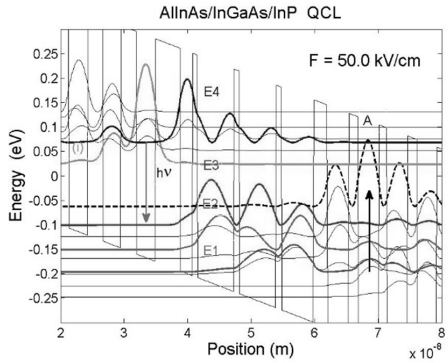

Fig. 13. The conduction band profile and squared wave-functions of moduli in the injector/active/injector segment of the laser under the applied field F = 50 kV/cm (at threshold). The wave-functions have been shifted to the energy positions of the respective levels. E4, E3, E2 and E1 refer to the upper, lower and ground states of lasing transitions. The black dashed line denotes the "spurious" excited state in the injector mini-gap. The lowest energy state in the injector couples directly to the upper laser level E4 [45].

The active region of the laser is embedded in the waveguide formed from the lower side by a low-doped InP substrate and from the upper side by a 2.5 μm AlInAs layer. The layer structure of AlInAs/InGaAs/InP laser is shown in Table 3. The whole laser structure was grown by MBE.

Table 3. A layer structure of AlInAs/InGaAs/InP lasers.

500 nm	InGaAs	n = 8e18 cm^{-3}	
2.5 μm	AlInAs	n = 1e17 cm^{-3}	Upper Waveguide
500 nm	InGaAs	n = 4e16 cm^{-3}	
~1.8 μm	**30 x AlInAs/InGaAs**		Active Region
500 nm	InGaAs	n = 4e16 cm^{-3}	Lower Waveguide
500 μm Substrate	InP	n = 2e17 cm^{-3}	

The lasers work at up to 340 K (~60ºC), emitting tens of mW of pulse power. At 20ºC the optical power per uncoated facet is of the order of 0.6 W. The slope efficiency is up to 1 W/A at room temperature; the wall plug efficiency is ~4%. The room-temperature light-current and current-voltage characteristics of the laser emitted at 9.2 μm are shown in Fig. 14.

The laser parameters can be further improved by optimizing the waveguide design, *i.e.*, by employing a symmetric InP waveguide. In this case, a conductive substrate can be used to grow the lower waveguide by MOVPE, then the active region is grown by MBE, and finally the upper waveguide is completed by MOVPE. That complicates technology, however a clear advantage, despite the increase of confinement factor, is the suppression of the free carrier absorption in the lower waveguide and consequent lowering of the threshold current. The parameters of developed devices are summarized in Table 4.

The lasers with a symmetric InP waveguide, due to a lower threshold current, should enable obtaining higher output powers which is advantageous in optical communication systems. Additionally, when processed into a buried active region type device, they can be CW-operated.

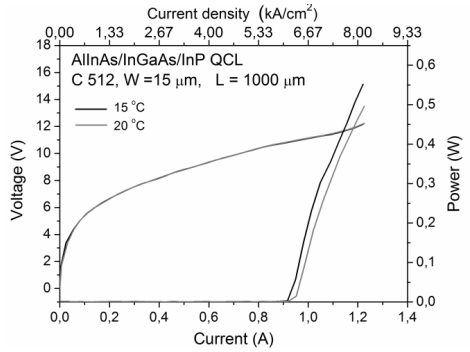

Fig. 14. Light-current and current-voltage characteristics of the Al0.48In0.52As/In0.53Ga0.47As/InP (λ = 9.2 μm) laser driven by 200 ns pulses with a repetition rate of 1 kHz [45].

Table 4. Parameters of AlInAs/InGaAs/InP lasers.

Parameter	Symbol	Value
Threshold current density	J_{th}(77 K)	2 kA/cm² – 3 kA/cm²
	J_{th}(300 K)	~5 kA/cm²
Threshold voltage	V_{th}(77 K)	10 V
	V_{th}(300 K)	11 V
Peak power (per facet)	P (77 K)	~2 W
	P (300 K)	~0.6 W
Max operating temperature	T_{max}	340 K
Characteristic temperature	T_0 (K)	120 K – 140 K
Differential gain	gΓ (77 K)	(5.7–6) cm^{-1}/kA
	gΓ (250 K)	(1.9–2.5) cm^{-1}/kA
Waveguide losses	α_w (77–300) K	~18 cm^{-1}
Wall-Plug efficiency	WPE	~4 %
Slope efficiency	η_{ext}	~1 W/A

8.3. MCT detection modules

VIGO System S.A. produces unique infrared detection modules (Fig. 15) that integrate in common packages infrared photodetectors, Peltier coolers, signal processing electronics, heat dissipation systems and other components [47]. The integration makes the detectors less vulnerable to electromagnetic interferences, over-bias, electrostatic discharges, and other environmental exposures. Additional advantages of the integration are: improved HF performance, standardization of the output signal, miniaturization and cost reduction.

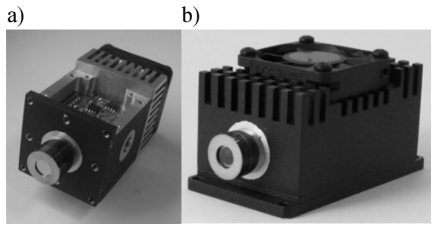

Fig. 15. Detection modules with a photodiode: unbiased (a) and biased (b).

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

The photodetector submodule (Fig. 16) consists of an optically immersed LWIR (~10 µm) hetero-structure photodiode, a 4-stage Peltier cooler and a temperature sensor housed in a hermetically sealed TO-8-based package, designed for detector operation at a temperature of approx. 200 K. The package is evacuated and then backfilled with a krypton/xenon mixture. It is supplied in optically transparent windows, absorbers of residual active gases (H_2O, O_2 and CO_2), convection and cold shields.

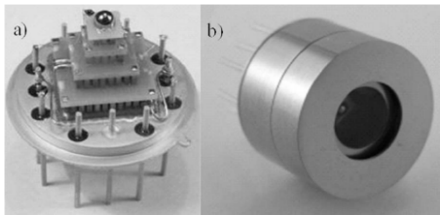

Fig. 16. An optically immersed detector mounted on a four-stage Peltier cooler (a) and a TO-8 header-based detector housing (b).

The photodiode construction is based on a modified HgCdTe PIN hetero-structure, grown by Metal Organic Vapor Phase Epitaxy [48–51]. Its architecture has been optimized by numerical simulation using a commercially available APSYS software package. The photodiode design is aimed to achieve fast and efficient collection of charge carriers, a low electric capacitance (of both depletion and diffusion layers) and a low series resistance [52].

The use of monolithic optical immersion of active elements in a high refraction index hyper-hemispherical lens results in a dramatic improvement of performance compared with the non-immersed device of the same optical size, namely a decrease of electric capacitance and dark current by two orders of magnitude [53]. In addition, the use of the double pass of radiation for enhanced absorption makes possible the reduction of the absorber thickness keeping unchanged the quantum efficiency.

A measured current-voltage plot of the photodiode (Fig. 17) shows three different bias ranges. The dark current initially increases with the reverse bias voltage, saturating in a range from −60 mV to −100 mV and then increases at higher voltages. A more close analysis reveals the diffusion nature of the dark current at low bias with a significant influence of series resistance at weak reverse bias. The dark current in the saturation range is mostly due to the Auger 7 and Shockley-Read-Hall thermal generation. The increase of dark current at a biasing voltage higher than −200 mV is due to the tunnel current, reduced to some degree by the series resistance.

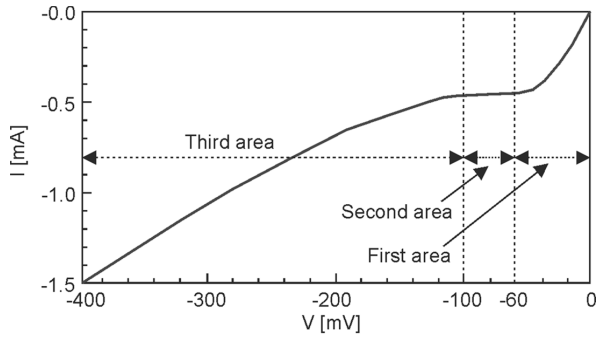

Fig. 17. A reverse dark current-voltage plot measured at 200 K. The background radiation was suppressed by 200 K cold shield. A physical photodiode area is $0.03 \times 0.03\text{mm}^2$.

Figure 18 shows a block diagram of the module electronic circuitry. The electronic circuitry provides optimized conditions for the photodiode operation – a constant voltage bias and a readout of current mode. The first stage is a DC-coupled trans-impedance preamplifier of a low input resistance, based on OPA 847 opamp, characterized by a low input noise voltage (0.85 nV/Hz$^{1/2}$) and a moderate input noise current (2.5 pA/Hz$^{1/2}$). The second stage is a ~20 dB AC-coupled voltage preamplifier with a 50 Ω output resistance, rejecting the DC phodiode signal component.

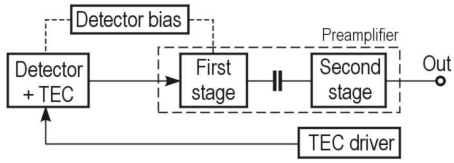

Fig. 18. A block diagram of the detection module.

A Peltier cooler driver has been used to keep temperature constant with accuracy better than 0.1ºC, at ambient temperatures of up to 50ºC. The heat generated inside the module by the Peltier cooler and the signal processing circuitry is dissipated with miniaturized fans.

Spectral responses of the detection modules have been measured using blackbody calibrated Fourier Transform Spectrophotometers. The module output noise voltage was determined with a signal analyser. Spectral detectivity was calculated as the signal-to-noise ratio, normalized to 1 cm^2 detector optical area and 1 Hz bandwidth (Fig. 19). The observed increase of detectivity with bias is mostly due to the elimination of recombination noise of the photodiode and the increased ratio of the series-to-parallel diode resistances, resulting in a significant increase of the current responsivity. It should be noted, that detectivity decreases at low (<~ 100 kHz) frequencies due to 1/*f* noise of the biased photodiodes. In contrast, the unbiased photodiodes are practically 1/*f* noise-free, similarly to other optical detectors [54].

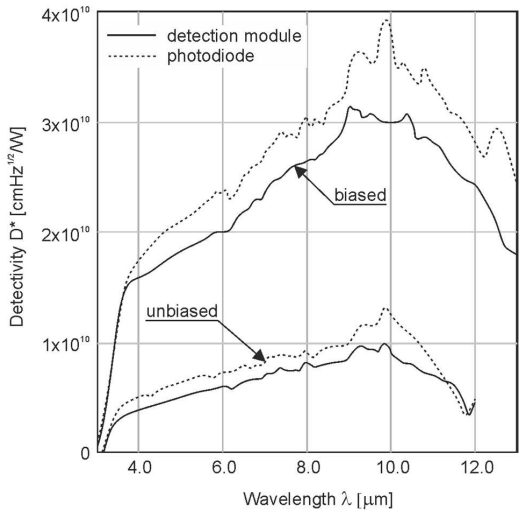

Fig. 19. The spectral detectivities of the photodiodes operating at zero and 200 mV bias measured for the detection module and for the photodiode itself (with subtracted preamplifier noise).

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 653–674.

A frequency response of the modules was measured using an optical parameter oscillator generating ~25 ps pulses of 10 µm infrared radiation. A signal decay time constant was measured by an 8 GHz bandwidth oscilloscope. Table 5 shows basic parameters of the two modules, with unbiased and 0.2 V biased ~10 µm photodiodes, respectively.

Table 5. Basic parameters of the detection modules.

Parameter	Unit	Module 1	Module 2
Photodiode area	mm	0.3 × 0.3	0.3 × 0.3
Photodiode temperature	K	200	200
Bias voltage	V	0	-0.2
Transimpedance @ R_{LOAD} = 50 Ω	V/A	$27 \cdot 10^3$	$13.5 \cdot 10^3$
Output resistance	Ω	1000	50
Gain bandwidth	MHz	0.001÷150	0.001÷1000
Noise voltage	$nV/Hz^{1/2}$	430	210
Voltage responsivity @ 10 µm	V/W	130000	210000
Detectivity @ 10 µm	$cmHz^{1/2}/W$	$9 \cdot 10^9$	$3 \cdot 10^{10}$
Time constant	ns	4.0	0.26

Basing on the measurement data it can be seen that the reverse biased modules could achieve detectivity smaller by a factor ~2 than the fundamental 300 K BLIP limit of performance (FOV = 180°). At present, intensive research is under way on miniaturized modules for >1 GHz bandwidth with detector and electronic blocks hermetically sealed in miniaturized packages (Fig. 20).

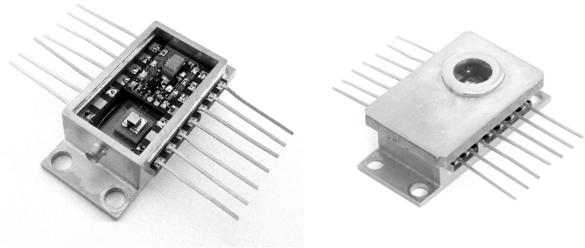

Fig. 20. Miniaturized detection modules before and after sealing.

9. Summary

The described analysis presents a high potential of FSO technology in the development of wireless mobile communication systems. The properties of FSO systems can be defined by both transmitter and receiver parameters. However, the greatest limitation of these systems is directly determined by the influence of the atmosphere. In this case, minimizing this negative effect mainly depends on the wavelength range of FSO link operation. The research performed so far has shown that a very promising solution may be the use of transceivers operating in the LWIR spectrum. This spectrum corresponds to the atmospheric transmission window while

being less affected by scattering and scintillation in comparison with existing NIR- or SWIR-FSO systems.

To further increase the availability of wireless communications, so-called hybrid links can be used. These constructions use both optical and radio communication channels. The preliminary research on this type of devices is currently underway. These studies are also the primary task of the described LasBITer project. In this project, a unique combination of free-space optical and radio links is used. The essential elements of the FSO link are quantum cascade lasers and MCT integrated detection modules. These lasers can be potentially used in other applications as well (*e.g.* Raman spectroscopy at different excitation wavelengths for advance chemical compounds detection [55]). The described processes of these devices' design have resulted in obtaining the optimal time-energy parameters of generated pulses, as well as the ability to detect ultra-low power LWIR radiation pulses .

Acknowledgements

This research was supported by The Polish National Centre for Research and Development grant DOB-BIO8/01/01/2016.

Refernces

[1] Taslakov, M.A., Simeonov, *et al.* (2004). Quantum cascade laser based system for line-of-sight data transmission in the mid IR. *Proc. SPIE 5830*, doi:10.1117/12.618774.

[2] Uysal, M., *et al.* (2016). Optical Wireless Communications An Emerging Technology. Signals and Communication Technology, 1–23.

[3] Saini, S., Gupta, A. (2014). Investigation to Find Optimal Modulation Format for Low Power Inter-Satellite Optical Wireless Communication. *Wireless and Optical Communications Networks (WOCN)*, doi: 10.1109/WOCN.2014.6923094.

[4] Alkholidi, A.G., Altowij, K. (2012). Optical Communications Systems: Effect of clear atmospheric turbulence on quality of free space optical communications in Western Asia. *InTECH*, doi 10.5772/1807.

[5] Fletcher, G.D.T., Hicks, R., Laurent, B. (2002). The SILEX optical interorbit link experiment. *IEEE J. Elec. & Comm. Eng.*, 3(6), 273–279.

[6] Muth, J. (2017). *Free-space Optical Communications: Building a 'deeper' understanding of underwater optical communications*. Laser Focus World.

[7] Wilson, K.E., Lesh, J.R. (1993). An overview of galileo optical experiment (GOPEX), Tech Report: TDA progress Report 42–114. *Communication Systems Research Section*, NASA.

[8] Wilson, K.E. (1996). An overview of the GOLD experiment between the ETS-VI satellite and the table mountain facility, TDA Progress Report 42–124. *Comm. Sys. and Research Sec.*, 9–19.

[9] Chlestil, Ch., *et al.* (2007). Optical wireless on swarm UAVs for high bit rate application. *The Mediterranean Journal of Computers and Networks*, 3(4), 142–150.

[10] Ortiz, G.G., *et al*. (2003). Design and development of a robust ATP subsystem for the altair UAV-to-ground lasercomm 2.5-Gbps demonstration. *Proc. SPIE 4975*, 103–114.

[11] *http://www.cyberbajt.pl/raport/40/0/142/*

[12] Wang, Ch.Y., *et al.* (2009). Mode-locked pulses from mid-infrared Quantum Cascade Lasers. *Optics Express*, 17(15), 12929–12943.

[13] Catalogue of Sonardyne firm (2017). Sonardyne Subsee Technology.

[14] Sadiku, M.N.O., *et al.* (2016). Free Space Optical Communications: An Overview. *European Scientific Journal*, 12(9).

[15] Kazaura, K., *et al.* (2008). Studies on next generation access technology using radio over free space optic links. *2nd International Conference on Next Generation Mobile Applications, Services, and Technologies – presentation*.

[16] Ramirez-Iniguez, R., Idrus, S.M., Sun, Z. (2007). *Optical Wireless Communications IR for Wireless Connectivity*. Taylor & Francis Group, CRC Press.

[17] Arnon, S. (2003). Optical Wireless Communications. *Encyclopedia of Optical Engineering.*

[18] Kasap, S., Ruda, H., Boucher, Y. (2009). *Cambridge illustrated handbook of optoelectronics and photonics*. Cambridge University Press.

[19] Bouchet, O., *et al*. (2010). *Free-Space Optics: Propagation and Communication*. Book, Wiley-ISTE.

[20] Talib, M.F., *et al.* (2017). Investigation on heavy precipitation effects over FSO link. *MATEC Web of Conferences*, 97, 01113 doi: 10.1051/matecconf/20179701113.

[21] Stull, R.B. (1988). *Atmospheric Sciences Library: An Introduction to Boundary Layer Meteorology*. Kluwer Academic Publishers.

[22] Alkholidi, A.G., Altowij, K.S. (2014). *Free Space Optical Communications –Theory and Practices*.

[23] Lawson, J.K., Carrano, C.J. (2006). Using Historic Models of Cn2 to predict r0 and regimes affected by atmospheric turbulence for horizontal, slant and topological paths. *Proc. SPIE 6303*, doi:10.1117/12.679108.

[24] Bloom, S. (2001). The physics of free-space optics. *AirFiber Inc*., 802-006-000, M-A1, 1–22.

[25] Singal, P., Rai, S., Punia, R., *et al*. (2015). Comparison of different transmitters using 1550 nm and 10 000 nm in FSO communication systems. *Int. Journal of Computer Science & Information Technology*, 7(3), 107–112.

[26] Berman, G.P., Chumak, A.A., *et al.* (2007). Beam wandering in the atmosphere: the effect of partial coherence. *Physical Review E*, 76, 056606-1–056606-7.

[27] Xian, Q., Wen-Yue, Z., *et al*. (2012). Long-distance propagation of pseudo-partially coherent Gaussian Schell-model beams in atmospheric turbulence. *Chin. Phys. B*, 2(9), 094202-1–094202-8.

[28] Zaki Rashed, A.N., Sharshar, H.A. (2014). Error Probability and Laser Beam Propagation Analysis in Local Area Optical Wireless Communication Networks Using Pulse Position Modulation Technique under Atmospheric Turbulence Effects. *International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)*, 3, 261–272.

[29] Willebrand, H., Ghuman, B. (2002). *Free Space Optics: Enabling Optical Connectivity in Today's Networks*. Sams Publishing.

[30] Bloom, S., Korevaar, E., *et al*. (2003). Understanding the performance of free-space optics. *Journal of Optical Networking*, 2(6), 178–200.

[31] Rongqing, H., O'Sullivan, M. (2009). *Fiber Optic Measurement Techniques*, 486–494.

[32] Forin, D.M., Incerti, G. (2010). Free Space Optical Technologies: *Trends in Telecommunications Technologies*, ed. Bouras, Ch.J.

[33] Altowij, K.S., Alkholidi, *et al*. (2010). The effect of Clear Atmospheric Turbulence on the Quality of the Free Space Optical Communications in Yemen. *Frontiers of Optoelectronics in China*, 3(4).

[34] Boone, B.G., Bruzzi, J.R., *et al*. (2004). Optical Communications Development for Spacecraft Applications. *Johns Hopkins Apl Technical Digest*, 25(4), 306–315.

[35] IEC 60825-1, International Standard, Safety of laser products, Edition 3.0 2014-05.

[36] Manor, H., Arnon, S. (2003). Performance of an optical wireless communication system as a function of wavelength. *Applied Optics*, 42(21), 4285–4294.

[37] Pavelchek, A., Trissel, R., *et al*. (2004). Long wave infrared (10 µm) Free Space Optical Communication. *Proc. of SPIE*, 5160, 247–252.

[38] Soni, G., Malhotra, J.T. (2011). Free Space Optics System: Performance and link availability. *International Journal of Computing and Corporate Research*, 1(4).

[39] Martini, R., Whittaker, E.A. (2005). Quantum cascade laser-based free space optical communications. *J. Opt. Fiber. Commun. Rep.*, 2, 1–14.

[40] Leitgeb, E., Plank, T., *et al*. (2014). Free Space Optics in different (civil and military) application scenarios in combination with other wireless technologies. *Telecommunications Network Strategy and Planning Symposium (Networks)*, doi: 10.1109/NETWKS.2014.6959207.

[41] Milner, S.D., Davis, C.C. (2004). Hybrid free space optical/RF networks for tactical operations. *Military Communications Conference (MILCOM)*, doi: 10.1109/MILCOM.2004.1493303.

[42] Akbulut. A., *et al.* An experimental hybrid FSO/RF communication system. Research supported by Ankara University Scientific Research Projects, Project No: 2001-00-00-006.

[43] Nadeem, F. *et al.* (2009). Weather effects on hybrid FSO/RF communication link. *IEEE Journal on Selected Areas in Communications,* 27(9).

[44] Faist, J. (2013). *Quantum cascade lasers*. Oxford University Press.

[45] Gutowski, P., Karbownik, P., *et al.* (2014). Room Temperature AlInAs/InGaAs/InP Quantum Cascade Lasers. *Photonics Letters of Poland*, 6(4), 142–144.

[46] Gutowski, P., Sankowska, I., *et al.* (2017). MBE Growth of Strain-Compensated InGaAs/InAlAs/InP Quantum Cascade Lasers. *Journal of Crystal Growth*, 466, 22–29.

[47] Gutowska, M., Gawron, W., *et al.* (2010). New Detection Modules for Free Space Optics. *Photonics Letters of Poland*, 2(2).

[48] Piotrowski, J., Orman, Z., *et al.* (2005). Uncooled long wave infrared photodetectors with optimized spectral response at selected spectral ranges. *Proc. SPIE*, 5783.

[49] Piotrowski, A., Gawron, W., *et al.* (2005). Improvements in MOCVD growth of Hg1-xCdxTe heterostructures for uncooled infrared photodetectors. *Proc. SPIE*, 5957, 108–116.

[50] Piotrowski, A., Klos, K., *et al.* (2007). Uncooled or minimally cooled 10μm photodetectors with subnanosecond response time. *Proc. SPIE*, 6542.

[51] Piotrowski, J., Rogalski, A. (2007). *High-Operating-Temperature Infrared Photodetectors*. SPIE.

[52] Piotrowski, J., Piotrowski, A. (2010). *Mercury Cadmium Telluride: Growth, Properties and Applications: Room temperature photodetectors*. ed. Capper, P., Garland, J., Willey.

[53] Piotrowski, J., Galus, W., *et al.* (1991). Near Room-Temperature IR Photo-detectors. *Infrared Phys*., 31, 11–48.

[54] Gnyba, M., Smulko, J., Kwiatkowski, A., Wierzba, P. (2011). Portable Raman spectrometer-design rules and applications. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 59(3), 325–329.

[55] Kwiatkowski, A., Czerwicka, M., Smulko, J., Stepnowski, P. (2014). Detection of denatonium benzoate (Bitrex) remnants in noncommercial alcoholic beverages by raman spectroscopy. *Journal of Forensic Sciences*, 59(5), 1358–1363.

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 675–683.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

TRANSMISSION QUALITY MEASUREMENTS IN DAB+ BROADCAST SYSTEM

Przemysław Gilski, Jacek Stefański

Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics, G. Narutowicza 11/12, 80-233 Gdańsk, Poland
(✉ pgilski@eti.pg.edu.pl, +48 58 347 6335, jstef@eti.pg.edu.pl)

Abstract

In the age of digital media, delivering broadcast content to customers at an acceptable level of quality is one of the most challenging tasks. The most important factor is the efficient use of available resources, including bandwidth. An appropriate way of managing the digital multiplex is essential for both the economic and technical issues. In this paper we describe transmission quality measurements in the DAB+ broadcast system. We provide a methodology for analysing parameters and factors related with the efficiency and reliability of a digital radio link. We describe a laboratory stand that can be used for transmission quality assessment on a regional and national level.

Keywords: broadcast technology, Digital Audio Broadcasting, Quality of Service, transmission quality, radio communication.

1. Introduction

Digital broadcasting systems, whether talking about DAB/DAB+ (*Digital Audio Broadcasting plus*) [1, 2] or other popular systems such as DMB (*Digital Multimedia Broadcasting*) [3] and DRM/DRM+ (*Digital Radio Mondiale*) [4], enable to transmit high quality speech and music signals compared with analogue AM (*Analog Modulation*) or FM (*Frequency Modulation*) radio. Aside from digital terrestrial services, there are also many satellite broadcasting systems operating worldwide [5, 6]. Furthermore, digital standards require less bandwidth per radio station, *i.e.* a single FM radio station requires a channel of 250 kHz, whereas about 12−15 radio programs in DAB+ require a channel of 1.5 MHz. This clearly shows that DAB+ is at least 2 times more bandwidth-efficient than FM.

The DAB+ standard is an evolution of the DAB standard, with a different source codec used for processing audio content. DAB+ uses an MPEG-4 (*Moving Picture Experts Group*) codec, compared with MPEG-2 used in DAB, which is more efficient and delivers higher quality at lower bitrates [7]. Additionally, all digital broadcasting standards enable to transmit additional information, including images and other interactive elements, *e.g.* EPG (*Electronic Program Guide*), well known from digital TV, programmable recording, *etc.* [8]. The digital standard has broad capabilities of regionalization, *i.e.* one nationwide service could be sacrificed in favor of a number of regional services.

Terrestrial broadcasting is the only free-to-air and cost-effective method for a truly mobile reception. However, broadcasters are not the same. They consist of public and private service broadcasters with a variety of national and regional stations. According to the EBU (*European Broadcast Union*), radio is: of vital cultural importance through Europe, consumed by a vast majority of Europeans every week, consumed at home, at work and on the move.

Conventional terrestrial radio transmission is faced with an increasingly strong competition from numerous streaming platforms and non-broadcast media, which use digital multimedia

Article history: received on Feb. 15, 2017; accepted on Jul. 31, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0050.

techniques to produce the optimum performance. Therefore, there is a growing demand for efficient ways of delivering high quality audio material at low bitrates, especially under bandwidth restrictions. This implies a necessity to monitor the transmission quality of offered services. Because the signal is affected by numerous factors in the propagation channel, it is necessary to control the quality of the broadcasted signal [9–11]. A case study concerning user expectations related with DAB+ can be found in [12].

2. DAB+ signal transmission

The DAB+ broadcasting system, based on OFDM (*Orthogonal Frequency-Division Multiplexing*) [13], can operate in a number of transmission modes, which define the number of parameters related to *e.g.* frame structure, subunits' quantity and length. The choice of a mode depends on system requirements and a type of transmission, *i.e.* terrestrial, satellite or hybrid, and carrier frequency. The DAB+ system can operate in 4 transmission modes:

1) Mode I – designed for terrestrial transmission in Band I (47–88 MHz), Band II (87.5–108 MHz) and Band III (174–240 MHz).
2) Mode II – used in terrestrial, satellite and hybrid transmission in L-Band (1452–1492 MHz).
3) Mode III – intended for terrestrial, satellite and hybrid transmission below 3 GHz.
4) Mode IV – applied similarly as Mode II.
5) The structure of a DAB+ frame consists of 3 elements, as shown in Fig. 1.

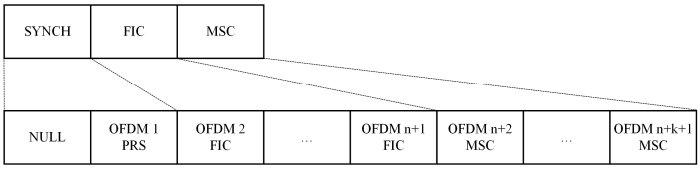

Fig. 1. A DAB+ frame structure.

The DAB+ frame comprises the following parts:

1) SYNCH (*Synchronization*) – responsible for synchronizing the transmitter and receiver, as well as frequency and gain adjustments.
2) FIC (*Fast Information Channel*) – transmits information about the configuration of the multiplex, including the number of services and assigned bitrate.
3) MSC (*Main Service Channel*) – contains the actual audio data.

The NULL symbol is used to determine the beginning of the DAB+ frame. If two successive symbols are known, the transmission mode can be easily determined on the receiver side. The PRS (*Phase Reference Symbol*) is the second OFDM symbol in the synchronization part. The receiver can also employ the PRS for more precise frame synchronization and frequency offset correction, which is accomplished by cross-correlation in time between the received and theoretical PRS.

The FIC contains information on how the multiplex is organized. Every receiver must process this data in order to present a list of available DAB+ services. The FIC consists of multiple FIBs (*Fast Information Blocks*), where each FIB contains 30 bytes of data and 16 bits of CRC (*Cyclic Redundancy Code*). Additional information concerning CRC algorithms can be found in [14]. Each FIB consists of multiple FIGs (*Fast Information Group*), which contain information about available services, their names and configuration.

The MSC is a time-interleaved data channel divided into a number of sub-channels, individually convolutionally coded, with EEP (*Equal Error Protection*) or UEP (*Unequal Error Protection*) error protection. Each sub-channel may carry one or more service components, *i.e.*

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 675–683.

audio or data, referred to as PAD (*Program Associated Data*). The information about sub-channel parameters is transmitted within CIFs (*Common Interleaved Frame*), as shown in Fig. 2.

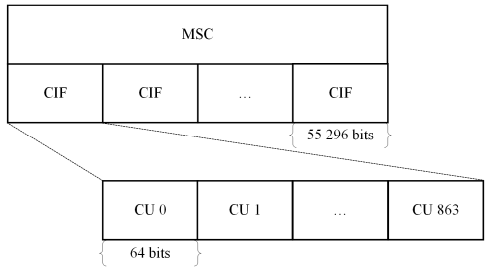

Fig. 2. An inner structure of MSC.

Each CIF comprises 55 296 bits, the smallest addressable unit is a CU (*Capacity Unit*), containing 64 bits. Therefore, a single CIF contains 864 CUs, addressed 0 to 863. Each CU may only be used for one sub-channel. After inserting additional correction mechanisms, an effective bitrate of the DAB+ stream is equal to 1152 kbps. Of course, a bitrate assigned to a particular service has a significant impact on quality perceived by the end user. Additional information may be found in [15].

3. Transmission quality measurements

Nowadays, mobile broadband networks carry multiple services that share radio access and core network resources. In addition, wireless networks must support delay-sensitive real-time services. Each service has different QoS (*Quality of Service*) requirements in terms of *e.g.* packet delay tolerance, acceptable packet loss rates and required minimum bit rates.

3.1. Quality of Service

The QoS parameter can be defined as a set of predefined technical specifications necessary to achieve the required service functionality. This can be an important factor when comparing services offered by different vendors or providers. When both price and feature are similar, quality becomes the key differentiator. Depending on the service being used, users have varying expectations concerning quality of performance and usability. Operators know, the better the experience, the longer and more frequently subscribers will consume content.

Quality plays a major role in wireless networks. Traffic management and optimization technologies enable network operators, as well as service providers and vendors, to improve subscriber QoS. As a result, it can help to attract new customers and raise their satisfaction. Additional information on quality measurements may be found in [16, 17].

3.2. Measurement stand

Broadcasting systems are capable of providing reliable digital services in real time to all users located in a predefined covered zone. One of the main factors is clearly the cost of an infrastructure and transmission power required to cover a given area. Another crucial aspect is the efficient way of monitoring transmission quality of offered services.

The measurement stand consists of a programmable receiver based on a DAB+ FM Digital Radio Development Board Pro platform for developing and evaluating DAB/DAB+,

SLideShow and FM with RDS (*Radio Data System*) services. It supports decoding multiple audio services, including DAB/DAB+ Band III and L-Band. The board contains a Keystone T2_L4A_8650C DAB/FM module and a Microchip PIC18F14K50 microcontroller. The device is powered by a USB Mini B connection, which is also used for communicating with the host computer. It has a 3.5 mm Stereo Jack connector for listening and a SMA (*SubMiniature version A*) connector for an external antenna [18]. A photo of the operating programmable receiver is shown in Fig. 3. Additional information on the design of a DAB/DAB+ receiver can be found in [19–21].

Fig. 3. Operating programmable DAB+ receiver.

The programmable receiver's GUI (*Graphical User Interface*) interface, written in C/C++, responsible for handling the device, is shown in Fig. 4.

Fig. 4. DAB+ receiver user interface.

The user interaction with the GUI is accomplished using a computer mouse and keyboard. The software has been designed to operate on any PC (*Personal Computer*) running Windows XP or higher.

3.3. Multiplex configuration

Today, one of the main objectives of national broadcasters and content providers is to design and implement viable services, which are based on new universal digital delivery systems. In Oct. 2016, the DAB+ multiplex in Gdańsk operated on channel 10D (215.072 MHz), transmission mode I. The configuration of the ensemble is described in Table 1. Each service had an EEP 3-A error correction and a coding efficiency of ½.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 675–683.

Table 1. DAB+ ensemble configuration in Gdańsk (Oct. 2016).

No.	Service	Bitrate [kbps]	No. of CU	Sub-channel	Start CU	Stop CU
1	PR Jedynka	112	84	1	0	83
2	PR Dwójka	128	96	2	84	179
3	PR Trójka	112	84	3	180	263
4	PR Czwórka	112	84	4	284	347
5	Radio Poland	64	48	5	348	395
6	Polskie Radio 24	64	48	6	396	443
7	Radio Rytm	96	72	7	444	515
8	Radio Gdańsk	104	79	8	516	593
9	Radio Dzieciom	72	54	9	594	647
10	Data	16	12	10	648	659
11	Journaline	16	12	11	660	671

According to the analysis, 672 CUs were allocated, whereas 192 CU remained free. All 11 services available in the ensemble occupied a total of 896 kbps, whereas 256 kbps remained unoccupied. The bandwidth occupancy of Band III (174–240 MHz) is shown in Fig. 5. This analysis was performed using an Anritsu Spectrum Master MS2724B [22].

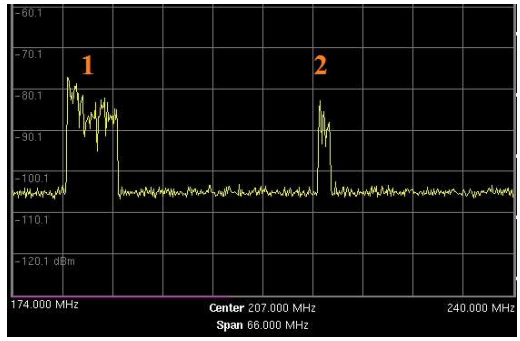

Fig. 5. Band III bandwidth occupancy.

As observed in Fig. 5, the signal at a centre frequency 184.5 MHz and a channel width 7 MHz (1) represents one of the DVB-T (*Digital Video Broadcasting – Terrestrial*) digital terrestrial television multiplex *MUX-8*, whereas the one at a centre frequency 215.072 MHz and a channel width 1.536 MHz (2) represents the DAB+ multiplex *DAB-GDA* [23].

3.4. Quality evaluation

In order to keep track whether the contracted QoS is being met, the parameters must be monitored and resources should be reallocated in response to system anomalies. If a change of state happens and the resource management cannot make resource adjustments to compensate it, the application can either adapt to the new level of QoS or to degrade to a reduced service level. The measurement of QoS is based on parameters including: delay, jitter, packet loss, throughput, SNR (*Signal-to-Noise Ratio*) and many other, depending on the application and management scheme. To ensure the transmission quality criterion, the DAB+ radio link was monitored over a period of 60 minutes during primetime, that is between 9 am and 10 am. The laboratory stand was set indoors. The structure of the DAB+ frame, as well as other parameters related with the standard, enable to monitor QoS in a number of ways. The operating parameters of the commercially available multiplex, concerning FIB Count, FIB CRC Errors, FIB Error Rate and SNR, are shown in Figs. 6–9.

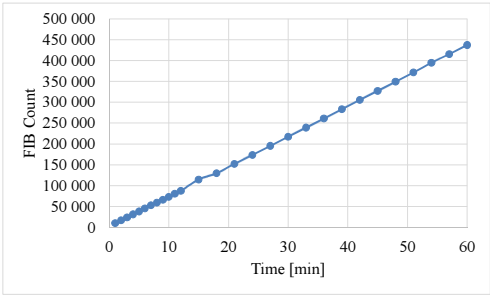

Fig. 6. Multiplex FIB Count.

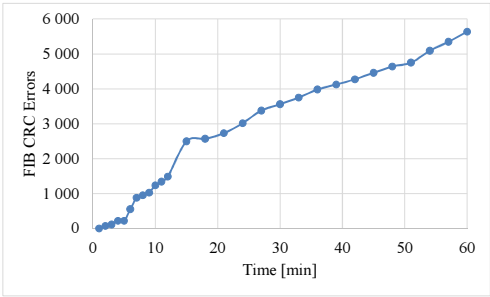

Fig. 7. Multiplex FIB CRC Errors.

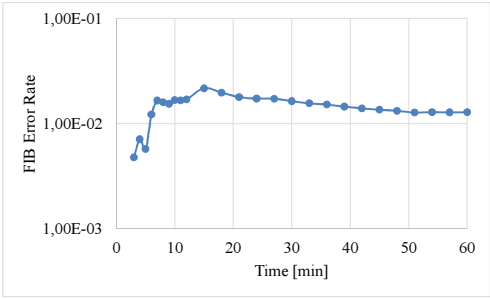

Fig. 8. Multiplex FIB Error Rate.

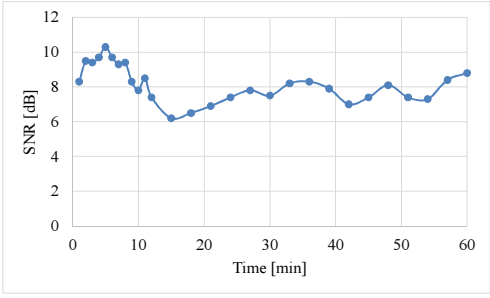

Fig. 9. Multiplex SNR.

As shown in Figs. 6–7, the commercial multiplex operates in an appropriate way. The character of both FIB Count, representing the total number of received FIB frames, defined as:

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 675–683.

$$FIB\_Count = \sum_{i=1}^{N} FIB_i , \tag{1}$$

and FIB CRC Errors, representing the total number of CRC errors, defined as:

$$FIB\_CRC\_Errors = \sum_{i=1}^{N} CRC\_Error_i , \tag{2}$$

is nearly linear. Otherwise, it would mean that a malfunction had appeared on the transmitter side. An FIB Error Rate – representing the relation between received erroneous and total FIB frames, defined as:

$$FIB\_Error\_Rate = \frac{N_E}{N_T} , \tag{3}$$

where N_E is the number of received erroneous FIB frames and N_T is the total number of received FIB frames – of $1.36 \cdot 10^{-2}$ was achieved by the multiplex. The average SNR oscillated around 8 dB.

Any malfunction on the transmitting side would lead to changes in the character of the FIG graphs, which are currently nearly linear. Additionally, any signal loss or occurring errors would lead to a decrease in SNR. When transferring these QoS parameters into QoE (*Quality of Experience*), of course this would be indicated by the user side, *i.e.* some services would be simply unavailable in particular conditions or time periods. Eventually, this would be clearly visible in lower values during assessment of the user quality. Currently, any interruption in offered services is viewed by the user as unacceptable.

If data continue to grow, broadcasters will be forced to manage quality in a more efficient way. The economic reality and physical limitations of available spectrum of resources prevent operators from simply adding more and more services. Broadcasters must plan today for future evolution of the network, which means working with parties that have a solid roadmap for QoS and transmission quality control mechanisms.

Additional to the transmission quality criterion, whenever planning a general or SFN (*Single Frequency Network*) DAB+ network, further studies on the electromagnetic compatibility, as well as compatibility with existing broadcasting services and networks should be carried out [24–27].

4. Conclusions

According to the report [28], as well as the European Radio Forum held in Kraków on Oct. 6th 2016 [29], a full introduction of DAB+ should be performed in cooperation with both the public and private broadcasters. Furthermore, as pointed out by the representatives of governments and the business sector, it should be preceded by further studies and research.

Moreover, national broadcasters of the Visegrad Group are planning to team up and introduce a new radio program, Radio V4. This program will be broadcasted in national languages of each country of the V4. It will include news and current affairs, as well as cultural, social and political topics [30]. This indicates that further work is required in order to provide reliable services at an acceptable level of quality. Additionally, transmission quality of offered services will have to be measured on regional, national and international levels.

As observed, the digital radio market continues to grow, and so does the demand for new efficient and reliable services. The proposed approach for monitoring transmission quality in the DAB+ broadcast system could be employed during the design and planning phases of a particular ensemble for both public and private broadcasters. Thanks to its portability and compatibility with Windows, it could be also used for evaluation and maintenance purposes on regional and nationwide levels.

Broadcasters, telecoms and content providers see the opportunity to offer more services, manufacturers look forward to selling larger quantities of devices and associated equipment. Network operators are keen on building new infrastructure. It is important to understand the pros and cons of different technologies and their commercial, economical and operational implications. Broadcasters will always aim to use the best possible means to reach the user. Users will welcome every new technology that offers more features and high quality content. When it comes to broadcasting, listeners are only interested in the quality, reliability and cost of a particular service.

References

[1] ETSI EN 300 401 European Standard. (2006). *Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers*. Sophia Antipolis Cedex, France.

[2] ETSI TS 102 563 Technical Specification. (2010). *Digital Audio Broadcasting (DAB); Transport of Advanced Audio Coding (AAC) audio*. Sophia Antipolis Cedex, France.

[3] Cho, S., Lee, G., Bae, B., Yang, K., Ahn, C.H., Lee, S.I., Ahn, C. (2007). System and Services of Terrestrial Digital Multimedia Broadcasting (T-DMB). *IEEE Transactions on Broadcasting*, 53, 171–178.

[4] ETSI ES 201 980 European Standard. (2014). *Digital Radio Mondiale (DRM); System Specification*. Sophia Antipolis Cedex, France.

[5] Kozamernik, F., Laflin, N., O'Leary, T. (2002). Satellite DSB systems – and their potential impact on the planning of terrestrial DAB services in Europe. *EBU Technical Review*, 1–17.

[6] Bem, D.J., Więckowski, T.W., Zieliński, R.J. (2000). Broadband satellite systems. *IEEE Communications Surveys & Tutorials*, 3(1), 2–15.

[7] Meltzer, S., Moser, G. (2006). MPEG-4 HE-AAC v2 – audio coding for today's digital media world. *EBU Technical Review*, 1–12.

[8] Kozamernik, F. (1995). Digital Audio Broadcasting – radio now and for the future. *EBU Technical Review*, Autumn, 2–27.

[9] Bosi, M., Goldberg, R.E. (2002). *Introduction to Digital Audio Coding and Standards*. Springer.

[10] Iwacz, G., Jajszczyk, A., Zajączkowski, M. (2008). *Multimedia Broadcasting and Multicasting in Mobile Networks*. John Wiley & Sons.

[11] Skarbek, W. (2016). *Foundations of Multimedia Techniques*. Warsaw University of Technology.

[12] Gilski, P., Stefański, J. (2016). Can the Digital Surpass the Analog: DAB+ Possibilities, Limitations and User Expectations. *International Journal of Electronics and Telecommunications*, 62(4), 353–361.

[13] Kowal, M., Kubal, S., Piotrowski, P., Zieliński, R.J. (2011). A Simulation Model of the Radio Frequency MIMO-OFDM System. *International Journal of Electronics and Telecommunications*, 57(3), 323–328.

[14] Peterson, W.W., Brown, D.T. (1961). Cyclic codes for error detection. *Proc. of the IRE*, 49, 228–235.

[15] Gandy, C. (2003). *DAB: an introduction to the Eureka DAB System and a guide to how it works*. BBC.

[16] Dymarski, P., Kula, S., Thanh, N. (2011). QoS Conditions for VoIP and VoD. *Journal of Telecom. and Information Technology*, 3, 29–37.

[17] Uhl, T., Paulsen, S. (2014). The new, parameterized VT Model for Determining Quality in the Video-telephony Service. *Bulletin of the Polish Academy of Sciences Technical Sciences*, 62(3), 431–437.

[18] Sixth Logic. (2012). *DAB+ FM Development Board User's Guide*.

[19] Taura, K., Tsujishita, M., Takeda, M., Kato, H., Ishida, M., Ishida, Y. (1996). A Digital Audio Broadcasting (DAB) Receiver. *IEEE Transactions on Consumer Electronics*, 42(3), 322–327.

[20] Cho, J., Cho, N., Bang, K., Park, M., Jun, H., Park, H., Hong, D. (2001). PC-Based Receiver for Eureka-147 Digital Audio Broadcasting. *IEEE Transactions on Broadcasting*, 47(2), 95–102.

[21] van de Laar, F., Philips, N., Huisken J. (1997). Towards the next generation of DAB receivers. *EBU Technical Review*, Summer, 46–59.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 675–683.

[22] Anritsu. (2012). *Spectrum Master MS2724B User Guide*.

[23] Digital terrestrial and satellite charts. (Oct. 2016). *http://www.sat-charts.eu*

[24] EBU. (2003). *Technical Bases for T-DAB Services Network Planning and Compatibility with Existing Broadcasting Services*. Switzerland.

[25] ETSI EN 302 077 European Standard. (2005). *Electromagnetic compatibility and Radio spectrum Matters (ERM); Transmitting equipment for the Terrestrial – Digital Audio Broadcasting (T-DAB) service*. Sophia Antipolis Cedex, France.

[26] Hunt, K.J., Cesky, T., Jeacock, T., Mägele, M., O'Leary, T., Petke, G. (1996). The CEPT T-DAB Planning Meeting. *EBU Technical Review*, Spring, 2–26.

[27] Brugger, R., Mayer, K. (2005). RRC-06 – technical basis and planning configurations for T-DAB and DVB-T. *EBU Technical Review*, 1–10.

[28] KRRiT. (2016). *Radio cyfrowe więcej niż radio: Zielona księga cyfryzacji radia w Polsce*. Warszawa.

[29] European Radio Forum. *http://www.polskieradio.pl/Europejskie-Forum-Radiowe/Tag179577* (Oct. 2016).

[30] Polish Radio. (Oct. 2016). *http://www.polskieradio.pl/7/129/Artykul/1670365,Wspolpraca-mediow-publicznych-grupy-V4-Barbara-Stanislawczyk-mamy-wspolne-korzenie*

www.czasopisma.pan.pl PAN www.journals.pan.pl

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 685–699.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

MINIMIZATION OF VENTILATOR-INDUCED LUNG INJURY IN ARDS PATIENTS – PART I: COMPLEX MODEL OF MECHANICALLY VENTILATED ARDS LUNGS

Jarosław Glapiński, Ireneusz Jabłoński

Wrocław University of Science and Technology, Faculty of Electronics, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
(✉ jaroslaw.glapinski@pwr.edu.pl, +48 71 320 6560, ireneusz.jablonski@pwr.edu.pl)

Abstract

A complex model of mechanically ventilated ARDS lungs is proposed in the paper. This analogue is based on a combination of four components that describe breathing mechanics: morphology, mechanical properties of surfactant, tissue and chest wall characteristics. Physical-mathematical formulas attained from experimental data have been translated into their electrical equivalents and implemented in MultiSim software.
To examine the adequacy of the forward model to the properties and behaviour of mechanically ventilated lungs in patients with ARDS symptoms, several computer simulations have been performed and reported in the paper. Inhomogeneous characteristics observed in the physical properties of ARDS lungs were mapped in a multi-lobe model and the measured outputs were compared with the data from physiological reports. In this way clinicians and scientists can obtain the knowledge on the moment of airway zone reopening/closure expressed as a function of pressure, volume or even time. In the paper, these trends were assessed for inhomogeneous distributions (proper for ARDS) of surfactant properties and airway geometry in consecutive lung lobes.
The proposed model enables monitoring of temporal alveolar dynamics in successive lobes as well as those occurring at a higher level of lung structure organization, *i.e.* in a point P_0 which can be used for collection of respiratory data during indirect management of recruitment/de-recruitment processes in ARDS lungs. The complex model and synthetic data generated for various parametrization scenarios make possible prospective studies on designing an indirect mode of alveolar zone management, *i.e.* with a minimized risk of repeated alveolar recruitment/de-recruitment and mechanical overstraining of lung tissues.

Keywords: lung alveolar surfactant, respiratory mechanics, mathematical modelling, medical decision support, lung protective ventilation.

1. Introduction

The clinical impact of *acute respiratory distress syndrome* (ARDS) and *acute lung injury* (ALI) is significant, with more than twenty thousand deaths per year [1]. There are reports which prove that improper mechanical ventilation of these patients can worsen ARDS mortality through a process of ventilator-induced lung injury [2]. What is more, it has been shown that lower tidal volumes and reduced plateau pressures led to a reduction of overall mortality [3]. However, a working point close to recruitment/de-recruitment phases brings a risk of repeated alveolar opening and collapse over time which contributes to further lung injury induced by ventilator [4–7].

One promoted strategy to prevent ventilator-induced lung injury is to open the lung and keep it open [8]. This scenario requires the recruitment manoeuver to open the lung, followed by a sufficient *positive end-expiratory pressure* (PEEP) to maintain the lung in its newly open state [9]. There are numerous theoretical and experimental works devoted to studies of pressure and time dependencies as the fundamental contributors to undesirable repeatability of recruitment/de-recruitment actions [10–14]. But there is a lack of modelling studies which explain relationships between geometry of alveolar zone, surfactant characteristics and ARDS

Article history: received on Jul. 10, 2017; accepted on Oct. 04, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0053.

lung mechanics expressed by pressure and flow trends recorded during mechanical ventilation, including consequences for optimal ventilation preserving repeated alveolar recruitment/de-recruitment.

Mathematical and physical-mathematical models of alveolar zone, which have been proposed in the literature [15], apply to respiratory mechanisms proper for breathing at volumes lower than the *total lung capacity* (TLC) [16–19]. The recruitment manoeuvre, especially in patients with ARDS or ALI (acute lung injury), is performed near and even above the TLC level. It means that the analogues reported in the literature are insufficient for reliable monitoring of alveolar conditions during recruitment of these structures over TLC. Additionally, logarithmic forms of models referenced in the literature [20] make impossible an easy description of peripheral lung mechanics at this volume level. Thus, replacing a logarithmic model with the polynomial one, which is well defined at/over TLC, and studying its properties and adequacy to the real conditions could be some kind of solution for this problem.

The paper presents a research program aimed at designing of a non-invasive procedure for monitoring of alveolar recruitment/de-recruitment based on pressure and flow signals measured at the mouth. Modelling of the respiratory system under mechanical ventilation is the first phase of this program. The proposed analogue, referenced to the experimental data, works near the TLC point and predicts mechanisms induced by surfactant properties during the recruitment manoeuvre in ARDS patients. The obtained results contribute to the understanding of complex interrelations occurring in ARDS lungs and signals generated in the model. Further, they can be used for designing an indirect measurement procedure for maintaining the working point of the lungs at the recruitment side of their pressure-volume characteristics, but close to the prevalent point located between the recruitment and the de-recruitment zones. Finally, minimization of the risk of repeated alveolar recruitment-de-recruitment and mechanical overstraining of lung tissues, at the same time, will result in limitation of lung injuries and their consequences, *i.e.* permanent and/or progressive limitation of the ventilation function.

2. Materials and methods

The proposed complex model of alveolar recruitment/de-recruitment is based on a combination of four elements that describe breathing mechanics: morphology, mechanical properties of surfactant, elasticity of tissue and the embedding of alveoli in the tissue (that is the connection with other alveoli via elastin and collagen fibres), and chest wall properties.

An alveolar pressure in static conditions can be described as the sum of pressures in consecutive structures of lung parenchyma, as it is shown in Fig. 1 and (1):

$$Palv = P_{\gamma} + Pt + Pcw + Pea, \tag{1}$$

where: $Palv$ – an alveolar airway pressure; $P\gamma$ – a surface tension pressure; Pt – a tissue pressure; Pcw – a rib cage and abdomen pressure; Pea – an external (ambient) pressure.

2.1. Modelling of alveolar geometry contribution in ventilated ARDS patients

A real shape of alveolar space in lung parenchyma is very complicated and unable to define [25]. Volume, thickness of alveoli wall, and microscopic anisotropy of expansion suggest a polyhedral shape of alveolus surface, but in many reports the shape of alveoli is assumed to be given by a simple geometrical model proposed to describe morphology properties [26, 27].

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 685–699.

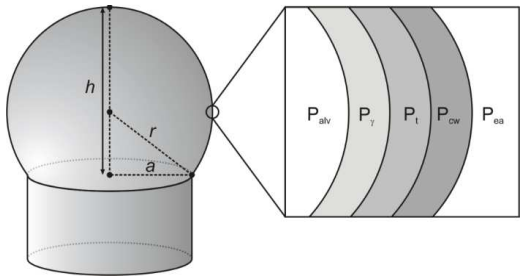

Fig. 1. Pressures defining equilibrium in the zone of alveolus: *a* – a diameter of the pulmonary bronchus; *r* – a diameter of the alveoli; *h* – a height of the alveoli.

For $P_{alv} = 0$ a rigid ring forms a flat surface at the alveolar mouth. During inflation the fluid-air interface of the alveolus becomes a part of spherical shape and forms a saucer to almost spherical shape [28]. Using the geometrical relations, a surface tension pressure of the surfactant can be calculated from the approximation of shape of the bubble by the Laplace's law:

$$P_\gamma = \frac{2\gamma}{r}, \tag{2}$$

where: γ – a surface tension; r – an alveolus diameter.

A volume of alveolus (V_A) can be calculated from the geometrical properties of a sphere shape:

$$V_A = \frac{1}{3}\pi h^2(3r - h), \tag{3}$$

$$r = \frac{a^2+h^2}{2h}.$$

If we assume a ring diameter (for a normal human lung being from 50 to150 μm) then changing the height of alveoli (h) we can obtain the pressure-volume characteristics for the described model, as shown in Fig. 2.

The presented relationship between the surface tension pressure and the alveolar volume can be estimated by a function which well corresponds to the calculated data:

$$P_\gamma = \gamma * \frac{V_A+k_1}{k_2*V_A^{k3}+k_4} + k_5. \tag{4}$$

Values of coefficients $k_1,\ldots,k_5$ estimated for several ring diameters are presented in Table 1.

2.2. Analogue of mechanical properties of surfactant

The lung surfactant properties contribute significantly to the temporal status of lung mechanics. The lipid mixture covering all alveoli reduces the surface tension of air-liquid interface. This decreases the alveolar pressure during inflation and in consequence the amount of energy required to open the lung. Disorder in the surfactant amount or properties dependent on pathological changes can cause global pulmonary diseases, like a respiratory distress. The surfactant properties measured with the captive bubble surfactometry were described by Lu *et al*. [29]. The axisymmetric drop shape analysis and captive bubble technique were proposed to measure the mechanic properties of bovine lipid extract surfactant related to the bubble surface area. The results of observations show that an increase in the surface area of the bubble implies an increase in the surface tension, like shown in Fig. 3, from its minimum (γ_{min}) to maximum value (γ_{max}) [20].

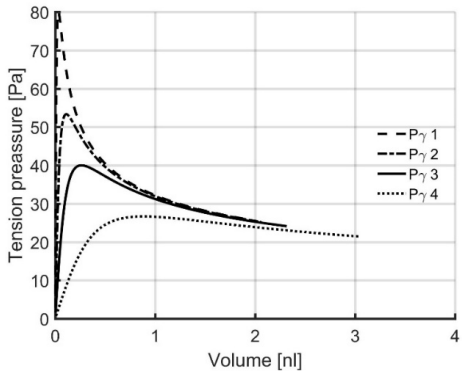

Fig. 2. Surface tension pressures ($P_\gamma i$) related to the alveolar volume for ring diameters $d_1 = 50$; $d_2 = 75$; $d_3 = 100$; $d_4 = 150$ and $\gamma = 1$.

Table 1. Values of coefficients estimated with (4) for selected diameters of alveoli.

	Diameter of alveolus (*r*)			
Parameter	50 μm	75 μm	100 μm	150 μm
k_1	–0.013	–0.073	–0.24	–1.2
k_2	0.036	0.033	0.029	0.021
k_3	1.5	1.5	1.6	1.7
k_4	0.0021	0.012	0.039	0.21
k_5	5.2	5.8	6.0	5.60

It has been reported that in healthy human lungs the surface tension of surfactant is about 2–23 mN/m, whereas pathological processes can increase its value up to 73 mN/m (*i.e.* up to the plasma tension) and reduce the hysteresis properties [27]. The experimental relation between changes in the surface tension and changes in the surface area of surfactant were reported in the literature [30] and approximated by the equations:

$$\gamma = aA_R^2 + bA_R^2 + c, \qquad A_R = \frac{A_A}{A_0}, \tag{5}$$

where: A_A – a relative surface area of alveolus at TLC; A_0 – an alveolar surface at a reference level, and estimated coefficients: a = 98,9 mN/m; b = –89.0 mN/m; c = 18.3 mN/m.

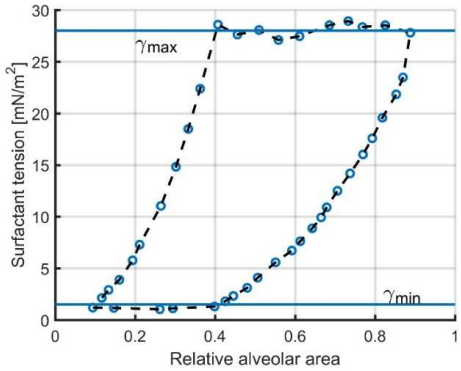

Fig. 3. The surfactant tension vs the relative alveolar area [29].

The hysteresis due to the properties of surfactant is caused by its structural transition from a mono- to multi-phase form and vice versa and is usually presented as the following function:

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 685–699.

$$A_0 = \begin{cases} A_A & \text{if } 1 < A_R \\ A_0 & \text{if } A_{Meta} < \mathrm{A_R} < 1 \\ A_A/A_{Meta} & \text{if } A_R < A_{Meta} \end{cases}, \tag{6}$$

where: A_{Meta} is a relative initial surface.

A relationship between the alveolar volume V_A and the surface area A_A was estimated by the polynomial function:

$$A_A = 0.56 * V_A^3 - 1.29 * V_A^2 + 1.73 * V_A^1. \tag{7}$$

Figure 4 shows relative changes of the alveolar area to TLC.

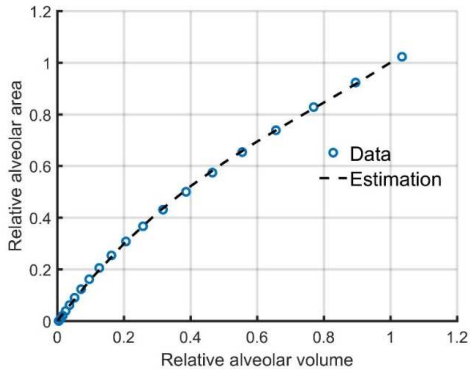

Fig. 4. The calculated dependence between relative changes in the alveolar area and in the alveolar volume.

2.3. Lung tissue model

Mechanical structures of lung parenchyma included in the model of lung tissue properties are described with the use of interpolation of the experimental results obtained during studies on animals [31]. This assessment is based on the equation by Andreassen [32]:

$$P_t = -ln\left(\frac{V_{A_{rel}}-1.01}{-1.31} + 0.1909\right) * 0.4320 \text{ kPa}, \tag{8}$$

where $V_{A_{rel}}$ is an alveolar volume at TLC.

In the paper, the model of lung tissue properties is estimated by the equation:

$$P_t = V_{A_{rel}}^{10} + 0{,}55 * V_{A_{rel}} \text{ kPa}, \tag{9}$$

which has the best correlation with the measured data [31], as in Fig. 5.

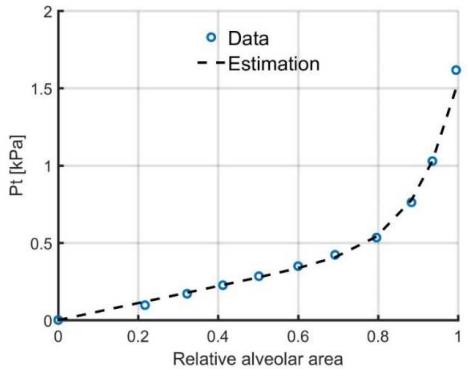

Fig. 5. A relationship between the elastic pressure and the relative volume of alveoli.

2.4. Chest wall model

A rib cage consists of curved ribs arranged in 12 pairs, ligaments and muscles, including the muscles between the ribs (intercostal muscles). These muscles on the one hand allow the ribcage to expand the lung while breathing, and to keep the negative intra-pleural pressure to prevent the alveoli to collapse while relaxation or sedation.

The chest wall properties were studied by Kimi Konno [33]. The oesophageal, gastric and mouth pressures were measured at different lung volumes during lung relaxation and with the closed mouthpiece. The presented data show the relationship between transthoracic and transabdominal pressures versus volume changes for standing and supine patient positions [34].

Steimle *et al.* proposed a mathematical description of the P_{cw} pressure dependent on the lung volume as the function [35]:

$$P_{cw} = 0.071 - ln\left(\frac{95\%}{(V_{Air}/V_{TLC}) - 22\%} - 1\right) * 0.58 \text{ kPa}. \tag{10}$$

This dependence was approximated by the following polynomial equation which enables modelling of the rib cage properties at volumes below 22% and over 100% of TLC (Fig. 6):

$$\begin{aligned} \mathrm{P_{cw}} = {} & 97{,}81 * \mathrm{V_{A_{rel}}^5} - 339{,}89 * \mathrm{V_{A_{rel}}^4} + 461{,}79 * \mathrm{V_{A_{rel}}^3} \\ & -306{,}13 * \mathrm{V_{A_{rel}}^2} + 101{,}56 * \mathrm{V_{A_{rel}}^1} - 13{,}59 \ \text{kPa}. \end{aligned} \tag{11}$$

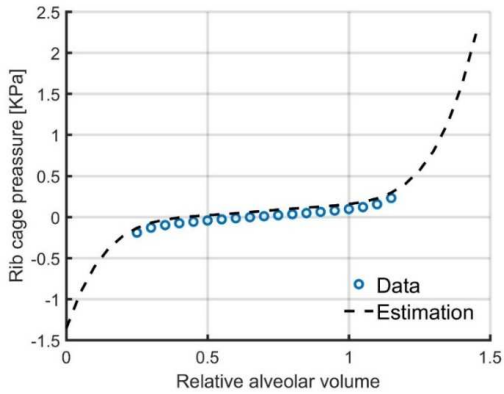

Fig. 6. A relationship between the rib cage pressure and the relative volume of alveoli.

2.5. Electrical model of lung recruitment in ARDS

The physical-mathematical model of ARDS lungs was translated into its electrical equivalent with the use of the electrical analogies and Multisim LabView software. An advantage of this approach and the programming tool is an easy implementation of electrical equivalents of physical systems, complex in structure and behaviour. What is more, for models described with a system of numerous differential equations the calculations are performed more efficiently in time [36, 37]. The proposed electrical equivalent suitable for imitation of lung recruitment/de-recruitment in a single lung lobe is presented in Fig. 7.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 685–699.

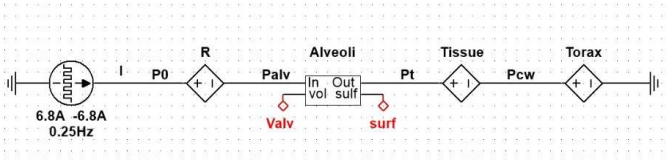

Fig. 7. An electrical forward model for simulation of lung recruitment/de-recruitment during ARDS in a single lobe.

I corresponds to an airflow source of 6,8 dm^3/s amplitude, thus a simulation of 1 s duration enables to show pressure/volume dependencies for volumes from 0 to 100% of TLC and from TLC to 0. *R* module represents an airway resistance pressure drop with its temporal value equal to $P_{res} = R \cdot I$, P_{alv} reconstructs an alveolar pressure drop as a function of volume (V_{alv}) and temporal surfactant tension (*surf*). The actual pressure (P_t) caused by tissue properties is modelled with (9) and changes of the thorax pressure (*Pcw*) are related to the thorax and abdomen properties, dependent on the lung volume (11).

The properties of surfactant hysteresis (Fig. 3) have been simplified to the alveolar area integration function (related to changes of the alveolar volume) with limits corresponded to the minimum (γ_{min}) and maximum (γ_{xax}) values (Fig. 8).

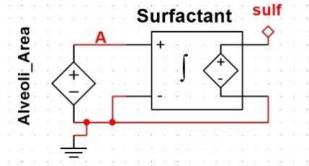

Fig. 8. An electrical equivalent of the simulation of mechanical properties of surfactant.

A challenge in management of the recruitment/de-recruitment process in ARDS lungs is the structural-parametric inhomogeneity. Thus, working with the analogues which implement this feature is fundamental for designing a reliable procedure of data processing during ARDS. The multi-compartment model of lungs from Fig. 9 addresses the defined demand. It enables to observe the pressure and flow signals at the mouth during recruitment/de-recruitment manoeuvers when different (*e.g.* inhomogeneous) parametric characteristics of lungs are implied in the successive branches.

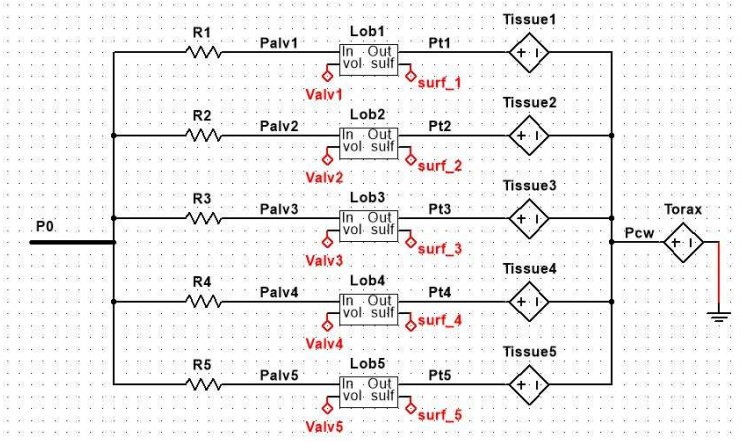

Fig. 9. A complex electrical forward model for simulation of lung recruitment/de-recruitment during ARDS.

In Fig. 9, each electrical branch simulates one lung lobe. TLC_i corresponds to the total capacity of i-th lobe, thus $TLC = \sum_{i=1}^{5} TLC_i$.

ARDS patients need the ventilation support in order to improve and stabilize the exchange of respiratory gas [37]. There are numerous scenarios designed for an optimal ventilation regime [38, 39]. All these works are aimed at efficient control of the recruitment mechanism and minimizing lung injury during the long-term mechanical ventilation [40, 41]. To prove feasibility of the minimal invasive, indirect management of ARDS lung recruitment, a model of respirator was proposed, as in Fig. 10, and integrated in MultiSim with the forward analogue from Fig. 9.

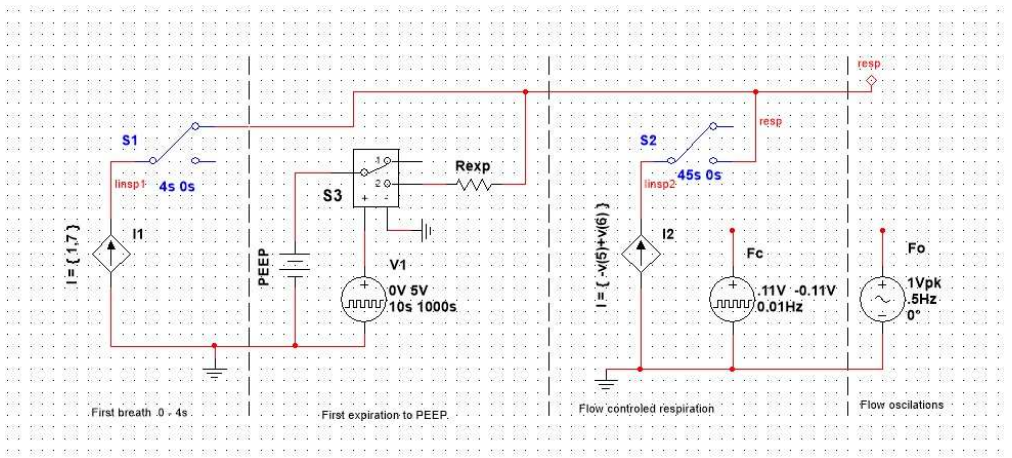

Fig. 10. An analogue of the respiratory signal of mechanical ventilation wave.

Its mode of operation assumes working as a generator of external excitation for the electrical equivalent of physiological system. Basically, a respiratory cycle comprises three phases in the designed model of respirator. Firstly, the lungs are filled with air in analogy to the first distending inspiration. To simulate this mechanism, the ventilation with a controlled flow (the airflow source $I1$) at a level of I_{insp1} = 1.7 l/s was applied for a duration of 4s in order to achieve a respiratory volume equal to TLC = 6.8 l. The spontaneous expiratory manoeuver up to the *positive end-expiratory pressure* (PEEP) level was performed in the second phase. Finally, the steady breathing (inspiration and expiration) was simulated in the third phase of work of the respirator. In this phase, the ventilation with a controlled airflow (I_{insp2} = 0.011 l/s) and a frequency (the pressure source F_c) equal to 0.01 Hz was imitated, but the model enables also switching to the forced oscillation mode simulated by a source of sinusoidal wave excitation (the pressure source F_o) generated with a frequency f_o = 0.5 Hz.

3. Results

To examine the adequacy of the forward model to the properties and behaviour of mechanically ventilated lungs in patients with ARDS symptoms, several computer simulations have been performed and reported as below.

First of all, a model for the constant flow respiratory scenario was studied. Since the first breath was taken from 0 to 100% of TLC, the changes of alveolar airway pressure, surface tension pressure, surfactant properties, tissue pressure and rib cage and abdomen pressure due to the volume evolution were assessed. The simulations were performed for the structure from Fig. 9 where the parametric description assumes cooperation of four pathological lobes and one

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 685–699.

physiological lobe in the respiratory system. More precisely, different diameters of alveolar rings (a = 50, 75, 100, 150 μm) and a homogenous surfactant tension γ = 73 mN/m were imitated in the pathological zone, whereas the parameter values in a single physiological lobe were set to a = 50 μm and γ = 2-23 mN/m, respectively. The observations from Fig. 11 are in accordance with the data from physiological reports which prove that a decrease in size of alveolar rings for an increased level of γ leads to increasing surfactant pressure tensions in subsequent compartments [42]. It means that a higher recruitment pressure is required to reopen the alveolar zone with that pathological characteristic.

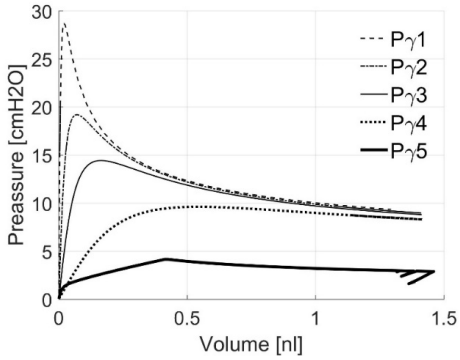

Fig. 11. The surfactant pressure tension related to the alveolar volume for lung lobes inhomogeneous in diameters (a) and surfactant tensions (γ).

An airway volume accessible for gas exchange in the alveolar space is also limited at the start of mechanical inflation of ARDS lungs (Fig. 12). In a healthy lung lobe V_{alv5} quickly inflates whereas for a pathological level of surfactant tension and the smallest alveolar ring a_1 = 50 μm the full capacity of alveoli V_{alv1} in the model of first lobe is available with a delay due to a larger shift in the recruitment moment. These shifts are clearly visible in Fig. 12 for consecutive lung lobes from the complex analogue – see trends for $V_{alv1} - V_{alv4}$.

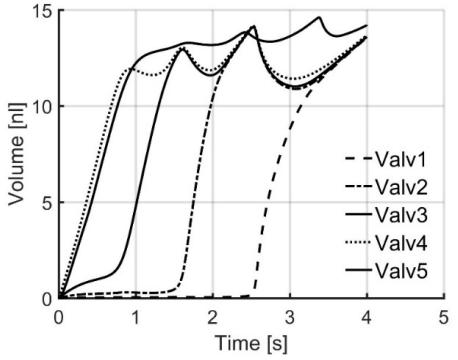

Fig. 12. Changes of volumes in successive lobes of the complex forward model of ARDS lungs.

For the clinicians involved in the management of lung recruitment/de-recruitment during the mechanical ventilation it is interesting to know the moment of airway zone reopening/closure expressed as a function of pressure, volume or even time [15]. Fig. 13 unveils this rule for alveolar pressures (P_{alv}) measured in healthy and pathological lung zones of the complex forward model of ARDS lungs proposed in Section 2 (Fig. 9). An advantage of the presented

modelling approach is that the user can test the contribution of various scenarios of heterogeneous parametrization of the surfactant and the airway geometry to the pressure and/or volume responses. For example, for a given γ = 73 mN/m, the pressure levels required to recruit the smallest three alveoli are the highest, and clear inflection points in trends P_{alv1} – P_{alv3}, apart from the moment of recruitment itself, show also the sensitivity of the model to the pendelluft between the compartments.

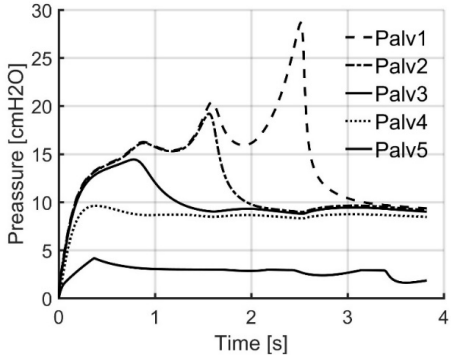

Fig. 13. Changes of pressures in successive lobes of the complex forward model of ARDS lungs.

A respiratory pressure (P_0) is presented in Fig. 14. This is a level of inference selected in the research protocol as the access point for measurement of pressure and flow signals which can be used for the inverse model identification, *i.e.* for non-invasive monitoring of lung recruitment/de-recruitment. In other words, information on the occurrence of recruitment/de-recruitment is extracted here by estimation of changes in the lung elasticity (compliance).

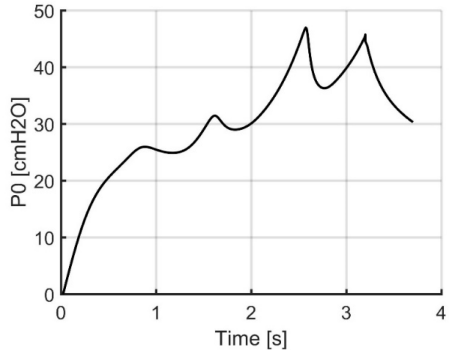

Fig. 14. Changes of pressure in the complex forward model of ARDS lungs.

The next stage of the forward model validation includes simulations of inspiration and expiration cycles in the analogue of mechanically ventilated ARDS lungs. In this phase of studies, the same value of airway ring diameter was applied in all lobes – $a_{1,\ldots,5}$ = 100 μm. The other parametrization was assumed in the forward equivalent to represent an inhomogeneous level of pathology in the surfactant properties, *i.e.* the values of surfactant tensions were equal to: γ_1 = 2–23 mN/m, γ_2 = 10–23 mN/m, γ_3 = 20–30 mN/m, γ_4 = 40–60 mN/m, γ_5 = 73–3 mN/m, respectively. Changes of alveolar volumes (P_{alv}) in successive lobes for reported conditions are shown in Fig. 15.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 685–699.

The observations are in agreement with the data from physiological reports, where for the homogeneous geometric conditions (the same diameters of airway rings) the availability of airway volume strongly depends on the surfactant tension γ [43, 44]. The higher γ, the higher the level of pathology, thus the greater shift in the moment of the airway recruitment expected. This is exactly the case proved in Fig. 15. What is more, the alveolar spaces in lobe 4 and lobe 5 are fully collapsed for some time during simulation.

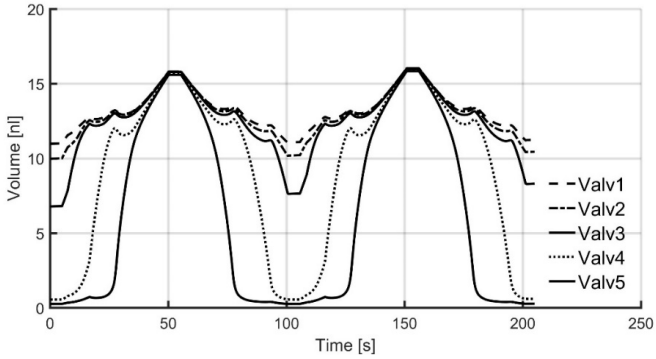

Fig. 15. Changes of volumes in successive lobes of the complex forward model of ARDS lungs during inspiration and expiration.

An inhomogeneous pathology in the surfactant tension (γ) changes the balance of the surface forces during the alveolar reopening/closure. This implies an increase in the level of alveolar pressure at what the alveoli in all modelled lobes start to recruit and stay open, regardless of either constant (Fig. 16) or sinusoid wave excitation applied during the mechanical ventilation (Fig. 17).

In fact, numerous studies have shown that ventilation with a variable wave of complex content is closer to the physiological conditions, resulting in the limitation of lung injury during the artificial ventilation action [45, 46].

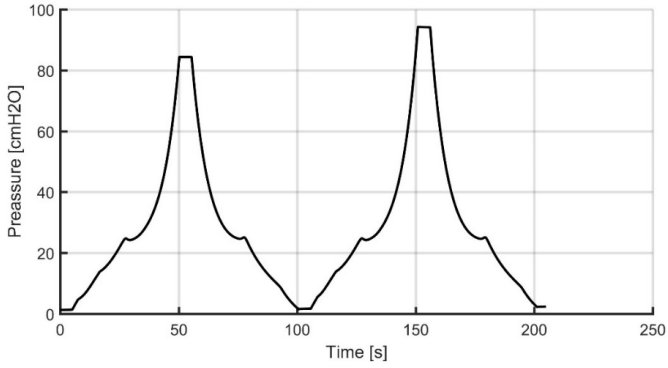

Fig. 16. A pressure response in the complex forward model of ARDS lungs during inspiration and expiration for a constant wave excitation.

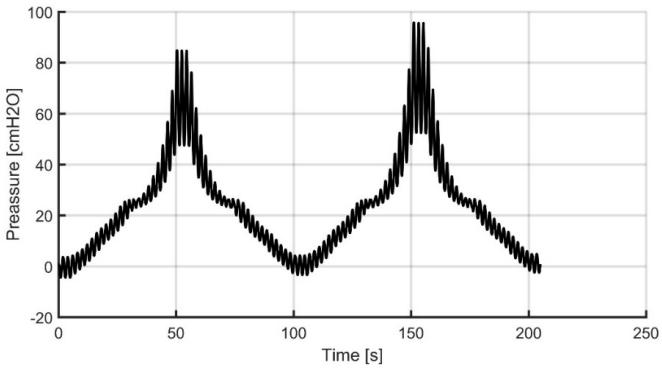

Fig. 17. A pressure response in the complex forward model of ARDS lungs during inspiration and expiration for a sinusoidal (f_{sinus} = 0.5 Hz) wave excitation.

4. Summary

The paper initiates the work focused on designing of a procedure for monitoring of alveolar recruitment/de-recruitment in ARDS lungs and automatized, clinical supporting the system in a stable and optimized recruitment regime. The optimization concerns here the identification of a working point where ARDS lungs do not move repeatedly between the recruitment and de-recruitment states and do not experience a significant overstraining due to the working far from TLC, at the same time. The main problem identified in this topic is a lack of the simulation and the inverse model(s) suitable for quantitative characterization of processes occurring at the alveolar level with the use of signals measured at the airway opening. In fact, these complex and inverse modelling can work in tandem during designing a data processing algorithm for ARDS lung management, *i.e.* a forward-inverse experiment can be organized to contribute to the knowledge about the rules governing the recruitment/de-recruitment processes and their crosstalk visible at a higher level of lung structure organization, *e.g.* in P_0 point (see Fig. 9). This is important since the clinical significance of the method requires non-invasive and thus indirect operation on a measured object. On the other hand, the clinical simplicity and usefulness of the final solution requires reliable operation on a limited amount of information and in the real-time mode.

The lung management of ARDS patients in clinical conditions has been carried out subjectively, *e.g.* based on observing characteristic levels in the output pressure and flow measured at the mouth and in the content of expiratory gas mixture. The primary goal is to identify and associate such characteristic features with physiological properties and to reconstruct the relations between the signals recorded at the airway opening and the processes measured in the alveolar space. Next, finding an algorithm of that mapping should be proposed in a real-time procedure for ARDS lungs.

A physical-mathematical forward model, including electrical equivalents, for the complex representation of processes and properties valid for ARDS lungs was proposed in the paper. This analogue includes a combination of four components that describe the breathing mechanics, *i.e.* ARDS lung morphology, mechanical properties of surfactant, tissue and chest wall characteristics. Here, especially the surfactant contribution to the observed output leads to the original inference into the temporal patterns of ARDS lungs' evolution during recruitment/de-recruitment. This was proved during the computer simulations for constant flow conditions and imitation of mechanical ventilation. It is possible to monitor pressure, flow and time dependencies during alveolar re-opening/closure processes using the complex model

described in the paper, both individually in distinguished lobes and at a level of their common output, *e.g.* a P_0 point. The model was validated qualitatively, referring to the trend properties and levels measured in the complex electrical equivalent and reported in the literature. The multi-lobe analogue is able to reveal a contribution of parametric heterogeneity to the outputs recorded at various levels of the proposed model. These facts can be exploited when creating an algorithm of the task of automatized, clinical management of recruitment processes during mechanically supported breathing, leading to the limitation of further lung injuries in ARDS lungs.

The next stages of scheduled research include designing of an inverse, time-varying parameter model of ARDS lungs, building a procedure of inverse model identification, and validating the proposed approach during computer simulations and trials with the use of an anesthetized pig. From this point of view it is sufficient that the forward model reconstructs the most significant properties and trends observed in a real object, including the orders of values of measured signals and parameters. The results presented in Section 3 validate positively the usefulness of the complex analogue for prospective research. The quantitative results of accuracy assessment of the applied approach are confirmed in the next stages of work.

In summary, the proposed complex model of mechanically ventilated ARDS is a prerequisite for an adequate inverse modelling [47] and then for designing a procedure of lung recruitment with a minimized injurious impact on lung tissues. These steps will be delivered in the next paper, together with an example of the experimental measurement proving correctness of the modelling studies and feasibility of the method. In detail, apart from the computer simulations, the program of prospective research will include experimental work with the use of a mechanically induced lung injury animal model, under general anaesthesia during the mechanical ventilation.

Acknowledgements

This work was supported by the National Science Centre under decision DEC-2013-11-B-ST7-01173.

References

[1] Rubenfeld, G.D. (2003). Epidemiology of acute lung injury. *Crit Care Med.*, 31, 276–284.

[2] Dreyfuss, D., Saumon, G. (1998). Ventilator-induced lung injury: lessons from experimental studies. *Am J. Resp. Crit. Care Med.*, 157, 294–323.

[3] Allen, G.B., Suratt, B.T., Rinaldi, L., Petty, J.M., Bates, J.H. (2006). Choosing the frequency of deep inflation in mice: balancing recruitment against ventilator-induced lung injury. *Am J. Physiol. Lung Cell Mol. Physiol.*, 291, L710–L717.

[4] Slutsky, A.S. (1999). Lung injury caused by mechanical ventilation. *Chest*, 116, 9–15.

[5] Hickling, K.G. (2001). Best compliance during a decremental, but not incremental, positive end-expiratory pressure trial is related to open-lung positive end-expiratory pressure: a mathematical model of acute respiratory distress syndrome lungs. *Am J. Respir. Crit. Care Med.*, 163, 69–78.

[6] Hickling, K.G. (1998). The pressure-volume curve is greatly modified by recruitment. A mathematical model of ARDS lungs. *Am J. Respir. Crit. Care Med.*, 158, 194–202.

[7] Pavone, L.A., Albert, S., Carney, D., Gatto, L.A., Halter, J.M., Nieman, G.F. (2007). Injurious mechanical ventilation in the normal lung causes a progressive pathologic change in dynamic alveolar mechanics. *Crit. Care*, 11, R64.

[8] Lachmann, B. (1992). Open up the lung and keep the lung open. *Intensive Care Med.*, 18, 319–321.

[9] Amato, M.B., Barbas, C.S., Medeiros, D.M., Magaldi, R.B., Schettino, G.P., Lorenzi-Filho, G., Kairalla, R.A., Deheinzelin, D., Munoz, C., Oliveira, R., Takagaki, T.Y., Carvalho, C.R. (1998). Effect of a protective-ventilation strategy on mortality in the acute respiratory distress syndrome. *N. Engl. J. Med.*, 338, 347–354.

[10] Alencar, A.M., Buldyrev, S.V., Majumdar, A., Stanley, H.E., Suki, B. (2001). Avalanche dynamics of crackle sound in the lung. *Phys. Rev. Lett.*, 87, 088101.

[11] Zhao, P., Yang, J., He, Y. (2017). Analysing the therapeutical action of lung recruitment maneuver on patients with acute respiratory distress syndrome by comparing different ventilation strategies. *Biomedical Research*, 28(4), 1828–1831.

[12] Bates, J.H., Irvin, C.G. (2002). Time dependence of recruitment and derecruitment in the lung: a theoretical model. *J. Appl. Physiol. Respir. Environ. Exercise Physiol.*, 93, 705–713.

[13] Brower, R.G., Morris, A., MacIntyre, N., Matthay, M.A., Hayden, D., Thompson, T., Clemmer, T., Lanken, P.N., Schoenfeld, D. (2003). Effects of recruitment maneuvers in patients with acute lung injury and acute respiratory distress syndrome ventilated with high positive end-expiratory pressure. *Crit. Care Med.*, 31, 2592–2597.

[14] Meade, M.O., Cook, D.J., Guyatt, G.H., Slutsky, A.S., Arabi, Y.M., Cooper, D.J., Davies, A.R., Hand, L.E., Zhou, Q., Thabane, L., Austin, P., Lapinsky, S., Baxter, A., Russell, J., Skrobik, Y., Ronco, J.J., Stewart, T.E. (2008). Ventilation strategy using low tidal volumes, recruitment maneuvers, and high positive end-expiratory pressure for acute lung injury and acute respiratory distress syndrome: a randomized controlled trial. *JAMA*, 299, 637–645.

[15] Albert, P., DiRocco, J., Allen, G.B., Bates, J.H.T., Lafollette, R., Kubiak, B.D., Fischer, J., Maroney, S., Nieman, G.F. (2009). The role of time and pressure on alveolar recruitment. *J. Appl. Physiol.*, 106, 757–765.

[16] Ma, B., Bates, J.H.T. (2010). Modeling the complex dynamics of derecruitmentin the lung. *Ann. Biomed. Eng.*, 38, 3466–3477.

[17] Bates, J.H.T., Irvin, C.G. (2002). Time dependence of recruitment and derecruitmentin the lung: A theoretical model. *J. Appl. Physiol.*, 93, 705–713.

[18] DiRocco, J.D., Carney, D.E., Nieman, G.F. (2007). Correlation between alveolar recruitment/derecruitment and inflection points on the pressurevolumecurve. *Intensive Care Med.*, 33, 1204–11.

[19] Hickling, K.G. (1998). The pressure-volume curve is greatly modified by recruitment. A mathematical model of ARDS lungs. Amer. *J. Respir. Crit.Care Med.*, 158, 194–202.

[20] Steimle, K.L., Mogensen, M.L., Karbing, D.S., de la Serna, J.B., Smith, B.W., Vacek, O., Andreassen, S. (2009). A Mathematical Physiological Model of the Pulmonary Ventilation. *Proc. of the 7th IFAC Symposium on Modelling and Control in Biomedical Systems Aalborg*, Denmark.

[21] Lutchen, K.R., Costa, K.D. (1990). Physiological interpretations based on lumped element models fit to respiratory impedance data: use of forward-inverse modeling. *IEEE Transactions on Biomedical Engineering*, 37, 1076–1086.

[22] Polak, A., Mroczka, J. (2006). Nonlinear model for mechanical ventilation of human lungs. *Comp. Biol. Med.*, 36(1), 41–58.

[23] Polak, A. (2002). A Morphometric Model of Lung Mechanics for Time-Domain Analysis of Alveolar Pressures during Mechanical Ventilation. *Ann. Biomed. Eng.*, 30(4), 537–45.

[24] Jabłoński, I., Mroczka, J. (2009). Frequency-domain identification of the respiratory system model during the interrupter experiment. *Measurement*, 42(3), 390–398.

[25] Valberg, P.A., Brain, J.D. (1977). Lung surface tension and air space dimensions from multiple pressure-volume curves. *J. Appl. Physiol.*, 43, 730–738.

[26] Sturm, R. (2015). A computer model for the simulation of nanoparticle deposition in the alveolar structures of the human lungs. *Annals of Translational Medicine*, 3(19).

[27] Reifenrath, R. (1975). The significance of alveolar geometry and surface tension in the respiratory mechanics of the lung. *Respiration Physiology*, 24(2), 115–137.

[28] Clements, J.A., Hustead, R.F., Johnson, R.P., Gribetz, I. (1961). Pulmonary surface tension and alveolar stability. *Journal of Applied Physiology*, 16(3), 444–450.

[29] Lu, J.Y., Distefano, J., Philips, K., Chen, A.W. (1999). Neumann Effect of the compression ratio on properties of lung surfactant (bovine lipid extract surfactant) films. *Respiration Physiology*, 115, 55–71.

[30] Sharp, J.T., Johnson, F.N., Goldberg, N.B., Van Lith, P. (1967). Hysteresis and stress adaptation in the human respiratory system. *J. Appl. Physiol*., 23(4), 487–97.

[31] Smith, J.C., Stamenovic, D. (1986). Surface forces in lungs. I. Alveolar surface tension-lung volume elationships. *J. Appl. Physiol*., 60, 1351–1350.

[32] Steimle, K.L., Mogensen, M.L., Karbing, D.S., Bernardino de la Serna, J., Andreassen, S. (2010). A model of ventilation of the healthy human lung. *Comput. Methods Programs Biomed*., 101(2), 144–155.

[33] Konno, K., Mead, J. (1968). Static volume–pressure characteristics of the rib cage and abdomen. *J. Appl. Physiol*., 24(4), 544–548.

[34] Goldman, M.D., Mead, J. (1973). Mechanical interaction between the diaphragm and rib cage. *Journal of Applied Physiology Published*, 35(2), 197–204.

[35] Steimle, K.L., Mogensen, M.L., Karbing, D.S., Bernardino de la Serna, J., Andreassen, S. (2010). A model of ventilation of the healthy human lung. *Comput Methods Programs Biomed*., 101(2), 144–55.

[36] Jabłoński, I., Polak, A.G., Mroczka, J. (2011). A preliminary study on the accuracy of respiratory input interrupter measurement using the interrupter technique. *Computer Methods & Programs in Biomedicine*, 101(2), 115–125.

[37] Jabłoński, I. (2013). Computer assessment of indirect insight during airflow interrupter maneuver of breathing. *Computer Methods & Programs in Biomedicine*, 110(3), 320–332.

[38] Kretschmer, J., Wahl, A., Moller, K. (2011). Dynamically generated models for medical decision support systems. *Computers in Biology and Medicine*, 41, 899–907.

[39] Malarkkan, N., Snook, N.J., Lumb, A.B. (2003). New aspects of ventilation in acute lung injury. *Anaesthesia*, 58(7), 627–728.

[40] Amato, M.B., Barbas, C.S., Medeiros, D.M., Magaldi, R.B., Schettino, G.P., Lorenzi-Filho, G., Kairalla, R.A., Deheinzelin, D., Munoz, C., Oliveira, R., Takagaki, T.Y., Carvalho, C.R. (1998). Effect of a protective-ventilation strategy on mortality in the acute respiratory distress syndrome. *N. Engl. J. Med.*, 338, 347–354.

[41] Putensen, C., Zech, S., Wrigge, H., Zinserling, J., Stuber, F., Spirgel, T., Mutz, N. (2001). Long-Term Effects of Spontaneous Breathing During Ventilatory Support in Patients with Acute Lung Injury. *Am J. of Resp. and Critical Care Med.*, 164(1).

[42] Harris, R.S.M. (2005). Pressure-Volume Curves of the Respiratory System. *Respiratory Care January*, 50(1).

[43] Schoel, W., Schürch, S., Goerke, J. (1994). The captive bubble method for the evaluation of pulmonary surfactant: surface tension, area, and volume calculations. *Biochimica et Biophysica Acta (BBA) – General Subjects*, 1200(3), 281–290.

[44] Schürch, S. (1982). Surface tension at low lung volumes: Dependence on time and alveolar size. *Respiration Physiology*, 48(3), 339–355.

[45] Boker, A., Haberman, C.J., Girling, L., Guzman, R.P., Louridas, G., Tanner, J.R., Cheang, M., Maycher, B.W., Bell, D.D., Doak, G.J. (2004). Variable ventilation improves perioperative lung function in patients undergoing abdominal aortic aneurysmectomy. *Anesthesiology*, 100(3), 608–616.

[46] Spieth, P., Carvalho, A., Pelosi, P., Hoehn, C., Meissner, C., Kasper, M., Hübler, M., von Neindorff, M., Dassow, C., Barrenschee, M., Uhlig, S., Koch, T., de Abreu, M.G. (2009). Variable Tidal Volumes Improve Lung Protective Ventilation Strategies in Experimental Lung Injury. *American Journal of Respiratory and Critical Care Medicine*, 179(8).

[47] Mroczka, J., Szczuczyński, D. (2009). Inverse problems formulated in terms of first-kind fredholm integral equations in indirect measurements. *Metrol. Meas. Syst*., 16(3), 333–357.

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

A FAST CLASSIFICATION METHOD OF FAULTS IN POWER ELECTRONIC CIRCUITS BASED ON SUPPORT VECTOR MACHINES

Jiang Cui, Ge Shi, Chunying Gong

Nanjing University of Aeronautics and Astronautics, College of Automation Engineering, 29 Jiangjun Road, Jiangning, Nanjing, Jiangsu, China (✉ pushriver@sina.com, +86 153 6600 6770, SG1503018@126.com, zjnjgcy@nuaa.edu.pl)

Abstract

Fault detection and location are important and front-end tasks in assuring the reliability of power electronic circuits. In essence, both tasks can be considered as the classification problem. This paper presents a fast fault classification method for power electronic circuits by using the *support vector machine* (SVM) as a classifier and the wavelet transform as a feature extraction technique. Using one-against-rest SVM and one-against-one SVM are two general approaches to fault classification in power electronic circuits. However, these methods have a high computational complexity, therefore in this design we employ a *directed acyclic graph* (DAG) SVM to implement the fault classification. The DAG SVM is close to the one-against-one SVM regarding its classification performance, but it is much faster. Moreover, in the presented approach, the DAG SVM is improved by introducing the method of *K*-nearest neighbours to reduce some computations, so that the classification time can be further reduced. A rectifier and an inverter are demonstrated to prove effectiveness of the presented design.

Keywords: power electronics, fault diagnosis, wavelet transforms, support vector machines, directed acyclic graph, nearest neighbours.

1. Introduction

Power electronic circuits (PECs) can be found extensively in industrial, military and residential applications [1]. High thermal and frequent mechanical stresses during the operations can accelerate the failure process of PECs. Once a fault inside a PEC occurs, unplanned electrical device breakdown may be triggered, in some cases associated with a very high cost or even casualties. For reasons of safety, reliability and maintenance, a fault has to be detected and diagnosed as soon as possible after its occurrence [2]. In some PECs with the fault tolerance capability, fault detection and diagnosis are necessary steps [3].

In essence, fault detection and fault diagnosis fall into the category of fault classification. Fault classification methods of PECs can be classified into three groups: model-based, expert system-based and *artificial intelligence* (AI)-based ones [4]. Among these methods, the AI-based ones seem to be attractive and interesting, because they have some advantages in comparison with other methods [1]. An AI-based method considers the whole system as a black box, whose inside details are being not relevant. This can avoid the problem of circuit system modelling. Classifiers based on the AI technique, such as the fuzzy inference method [5–7], have been proved to be effective. Compared with methods based on pure hardware circuits [8], AI-based methods usually employ algorithms, which can be easy and flexible to transplant and upgrade while keeping unchanged the corresponding hardware .

In the AI applications, the *Neural Network* (NN) is a good classifier, which has good performance in fault classification of PEC. In [9], the *radial-basis-function* (RBF) NN is adopted to perform fault detection of an induction motor drive circuit. A *back-propagation NN* (BPNN) is presented in [10] to diagnose faults of a three-phase inverter. In this example,

Article history: received on Nov. 12, 2016; accepted on May 21, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0056.

the classification accuracy of over 95% is reported. In [11], a multi-layered perceptron network is employed to classify open faults of a simulated *voltage source inverter* (VSI). Another study of BPNN application to a multi-level inverter fault classification is described in [2]. In some studies, two or more NNs are integrated to perform the classification task, which can improve the classification performance of a diagnostic system [12−14]. The NN-based method has also some drawbacks. For example, different NN trainings may lead to different classification results. In addition, high-dimensional data can result in a long training process, or even a convergence failure. Focusing on these drawbacks, the NN classifier can be improved with other methods. For example, in [15], before being input to the classifier, the fault samples are pre-processed with *Principal Component Analysis* (PCA) and *Genetic Algorithm* (GA), which can reduce dimensions of the training samples. In [16], the BPNN structure is optimized to improve its classification performance.

Recently, the applications of *Support Vector Machine* (SVM) to fault classification of PECs have been reported. The SVM has some excellent characteristics, *e.g.* it needs less adjustable parameters and can find the global solution easily during training, thus leading to stable classification results. The conventional SVM can create binary classes, and such a classifier is called a *binary SVM* (BSVM). In [17], one BSVM is used to detect whether the inverter is faulty, and the other BSVM can localize the faulty power switch (upper or bottom half-bridge). Another application of BSVM to the fault detection of an induction motor drive is described in [18]. Generally, diagnosing a PEC involves a multi-class classification. In the domain of machine learning, a multi-class classifier design for SVM involves two methods [19]. The first method is meant to create a multi-class classification in one step, whereas the second one – to combine several BSVMs to form a multi-class classifier, which has three basic forms: one-against-rest SVM, one-against-one SVM and *Directed Acyclic Graph* (DAG) SVM. In diagnosing a PEC, the one-against-rest SVM and one-against-one SVM have been used. For instance, in [20] and [21], the one-against-rest SVM classifiers are adopted to perform fault diagnosis of simulated rectifiers. Two examples of diagnosing inverters with one-against-rest SVMs are described in [22] and [23]. The application of one-against-one SVM to the diagnosis of an induction motor drive can be found in [24]. The use of DAG SVM, however, is seldom reported in fault classification of PECs.

According to the experiment results obtained in [19], both DAG SVM and one-against-one SVM are suitable for practical use, because they can always achieve high accuracy in applications. A one-against-one SVM creates classification with a greater number of computations, so that fault classification is a time-consuming task for this classifier. In this research, we apply a DAG SVM to PEC fault classification and, moreover, we attempt to improve the DAG SVM by employing an additional method. Compared with the conventional DAG SVM and other SVM classifiers, the new method needs less BSVMs to perform fault classification and, accordingly, the classification time can be shortened. Also, the classification accuracy of the presented method is very close to that of the conventional DAG SVM. Hence, the presented method can be considered as an alternative classifier for the DAG SVM. Experiments on a rectifier and an inverter were performed to prove effectiveness of the presented method. For the purpose of comparison, five classifiers were designed and examined regarding their classification accuracy and testing time.

2. Basic theories cocerning SVM classifier

2.1. Support vector machines for binary classification

A standard support vector machine classifier, invented by Vapnik and his colleagues [25], has a theoretical background of statistical learning theory and executes *Structural Risk*

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

Minimization (SRM) [26]. It can create a binary classification with excellent performance. The standard binary classifier can create both linear and nonlinear classifications. In the domain of fault detection and diagnosis of PECs, the nonlinear classification seems to be more practical. The nonlinear BSVM adopts a mapping function $\psi(\cdot)$, which can map data samples from the measurement space to a high-dimensional space. The binary classes can become linearly separable in the high-dimensional space. This principle is expressed in Fig. 1, where ○ and ● represent **class I** and **class II**, respectively.

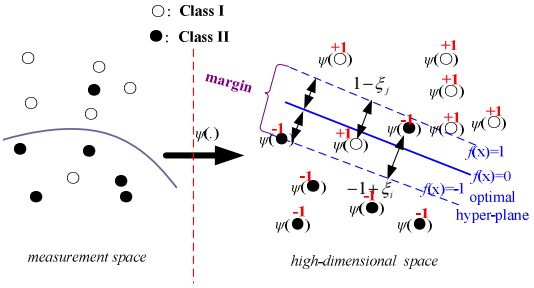

Fig. 1. A basic model of nonlinear BSVM.

In order to implement the BSVM, a margin between samples in the high-dimensional space should be maximized. Assume a data set to be $\{\boldsymbol{x}_i\}$ (i = 1, 2, …, Q, where Q is the number of data samples; $\boldsymbol{x}_i$ is an ith data sample in the measurement space), $\boldsymbol{x}_i \in R^d$ (R being the d-dimensional measurement space). After being mapped to the high-dimensional space with an inexplicit mapping function $\psi(\cdot)$, $\{\boldsymbol{x}_i\}$ turns into $\{\psi(\boldsymbol{x}_i)\}$. Then, each sample $\psi(\boldsymbol{x}_i)$ is assigned to a label y_i (y_i =+1 for **Class I**, and y_i=-1 for **Class II**). The optimal hyper-plane can be represented with:

$$f(\boldsymbol{x}_i) = \psi(\boldsymbol{x}_i)\cdot\boldsymbol{w} + b = 0\,, \tag{1}$$

where: $\boldsymbol{w}$ is a weight vector of optimal hyper-plane; b is a bias.

In order to allow some samples to be misclassified to reduce the effect on the decision boundary position, the slack variables $\xi_i \geq 0$ are necessary. Hence, by considering the misclassified samples, the samples try to be classified correctly beyond the margin:

$$y_i(\psi(\boldsymbol{x}_i)\cdot\boldsymbol{w} + b) - 1 + \xi_i \geq 0\,. \tag{2}$$

Maximizing the margin means minimizing the following quadratic optimization problem:

$$\begin{aligned} &\varphi(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 \big/ 2 + C\sum_{i=1}^{Q}\xi_i \\ &s.t. \qquad y_i(\psi(\boldsymbol{x}_i)\cdot\boldsymbol{w} + b) - 1 + \xi_i \geq 0 \end{aligned}\,, \tag{3}$$

where C is a penalty parameter for balancing the classification accuracy and complexity of the decision boundary.

Solving this optimization equation needs the *Lagrange multipliers* (LMs) $\lambda_i \geq 0$:

$$L = \varphi(\boldsymbol{w}) - \sum_{i=1}^{Q}\lambda_i(y_i(\psi(\boldsymbol{x}_i)\cdot\boldsymbol{w} + b) - 1 + \xi_i)\,. \tag{4}$$

By removing primal variables, the partial derivatives of L in respect to $\boldsymbol{w}$ and b are used to yield the dual formulation L*:

$$L^* = \sum_{i=1}^{Q} \lambda_i - \frac{1}{2} \sum_{i=1}^{Q} \sum_{j=1}^{Q} \lambda_i \lambda_j y_i y_j (\psi(\boldsymbol{x}_i) \cdot \psi(\boldsymbol{x}_j)^T), \tag{5}$$

where: λ_i, λ_j are LMs of *i*th and *j*th data samples, respectively; $(\cdot)$ is an inner product; *T* is the transpose of vector.

The solution of this optimization problem will generate *support vectors* (SVs), whose corresponding LMs are $\lambda_i > 0$. Let the number of SVs be n_{sv} and considering a kernel function $k(\boldsymbol{t}, \boldsymbol{x}_k) = (\psi(\boldsymbol{t}) \cdot \psi(\boldsymbol{x}_k)^T)$, the calculation function of BSVM becomes:

$$f(\boldsymbol{t}) = \sum_{k=1}^{n_{sv}} y_k \hat{\lambda}_k k(\boldsymbol{t}, \boldsymbol{x}_k) + b, \tag{6}$$

where: $\boldsymbol{t}$ is a data sample to be classified; $\hat{\lambda}_k > 0$ is an LM of *k*th SV $\boldsymbol{x}_k$.

The kernel function has several forms [27]. In our experiments, the RBF kernel function $(k(\boldsymbol{t}, \boldsymbol{x}_i) = \exp(-|\boldsymbol{t} - \boldsymbol{x}_i|^2 / \sigma^2)$, where $\sigma > 0$ is a kernel parameter; $\boldsymbol{x}_i$ is an *i*th SV) is considered, because this nonlinear kernel function can always lead to good classification performance.

2.2. Two conventional multi-class SVMs for PEC fault classification

A conventional one-against-rest SVM employs the *Winner-Takes-All* (WTA) rule to implement a pattern classification [28]. This classifier is simple to use in the fault classification of PEC. For *N* fault classes, *N* BSVMs are needed. For each training, an i^{th} class (labelled with "–1") is separated from the rest (*N*–1) classes (labelled with "+1"). Finally, a sample $\boldsymbol{x}$ should be assigned to the fault class whose corresponding decision function has the minimum value:

$$\arg \min_{i=1,2,\ldots,N} [f_i(\boldsymbol{t})], \tag{7}$$

where $f_i(\boldsymbol{t})$ is a decision function of *i*th BSVM for a sample $\boldsymbol{t}$.

For the one-against-one SVM, altogether $N(N-1)/2$ BSVMs are constructed. The decision function for the BSVM, which is formed by classes *i* and *j* ($i \neq j$), can be expressed in the form of:

$$f_{ij}(\boldsymbol{t}) = \sum_{k=1}^{n_{sv}^{i,j}} y_k^{i,j} \lambda_k^{i,j} K(\boldsymbol{t}, \boldsymbol{x}_k^{i,j}) + b_{i,j}^*, \tag{8}$$

where: $n_{sv}^{i,j}$ is the number of SVs; $\lambda_k^{i,j}$ is an LM of *k*th SV; $y_k^{i,j}$ is a label of *k*th SV; $b_{i,j}^*$ is a bias of this BSVM.

In the final stage, all decision functions of BSVMs need to vote for the appropriate class. The max-wins strategy is adopted to find a class which wins the maximum votes. However, with the increase of *N*, the number of computations for this classifier will increase drastically, and thus this method will probably become unsuitable for fast fault classification of PECs.

3. Presented SVM classifier

3.1. Typical Structure of DAG SVM

The classifier employed in our research is a typical DAG SVM [29], whose training phase is the same as in the one-against-one SVM by constructing $(N-1)N/2$ BSVMs for *N* fault classes. In the classification phase, the BSVMs are arranged to form a directed acyclic graph. Each node of the DAG SVM is represented with B*ij*, indicating a BSVM classifier corresponding to class *i* and class *j* ($i \neq j$). Except for the root node, each node has one input and two outputs which stand for the possible decision values (left and right branch) of the

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

BSVM. Fig. 2 shows a typical model of DAG SVM with $N = 5$ (for simplicity, assume five fault classes to be marked: '0', '1', '2', '3' and '4', respectively).

The DAG SVM classifier structure is similar to a pyramid, which can be partitioned into N layers for N fault classes. For instance, in Fig. 2, the root node is the first layer (containing B04), and the second layer contains two nodes (containing B03 and B14), … , the kth layer ($k < N$) contains k BSVMs, …, and so on. The final layer contains only leaves, which represent the five separated classes.

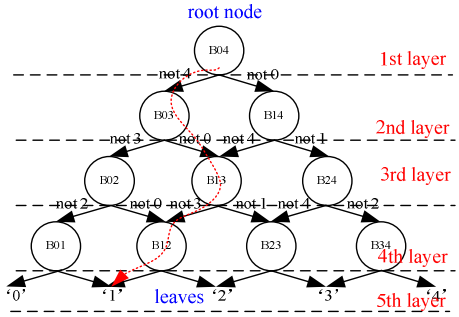

Fig. 2. DAG SVM Structure for $N = 5$.

The classification task is initiated from the first layer and is stopped in the final layer. In each layer (except for the final layer), one BSVM is evaluated to generate the output result, which becomes the input of a BSVM in the next layer. Fig. 2 gives an illustration of such a flow path (in red dashed curve), in which B04→B03→B13→B12→'1' are evaluated one after the other. Hence, for N fault classes, (N–1) BSVMs need to be evaluated for each classification task.

3.2. Improvement of DAG SVM

Generally, the DAG SVM is a fast classifier, but it still needs to compute (N–1) decision functions for the BSVMs. A typical DAG SVM always starts from the root node, however in this study we consider changing the starting node of DAG SVM, which will probably reduce the evaluation time of this SVM classifier. For example, in Fig. 2, if the starting node is initiated from B03, not from the root node, the evaluation flow path will become B03→B13→B12→'1'. In this case, the computation of BSVM decision function at B04 node can be bypassed. The key problem is, how to know it is the node B04 that should be avoided. In other words, how to find a limited set of nodes which participate in the computation.

In the paper there is adopted the method of *K-Nearest Neighbours* (K-NN) as an auxiliary classifier to find the limited set of nodes. The K-NN classifier is an easy to use, non-parametric method and. It was applied to the fault diagnosis of a generator rotor [30].

Assume N fault classes, each fault class containing L data samples. The centroid for class j is defined in the measurement space:

$$\boldsymbol{C}_j = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{x}_{ij}\,, \tag{9}$$

where $\boldsymbol{x}_{ij}$ is an ith training sample of class j ($j = 1, 2, \ldots, N$).

The K-NN method selects K closest neighbours basing on Euclidean distances between an unknown sample $\boldsymbol{x}$ and the centroids. K fault classes corresponding to the K closest centroids fall into a limited set, from which the starting node can be chosen. Fig. 3 illustrates the way of finding the starting node ($N = 5$) for $K = 3$.

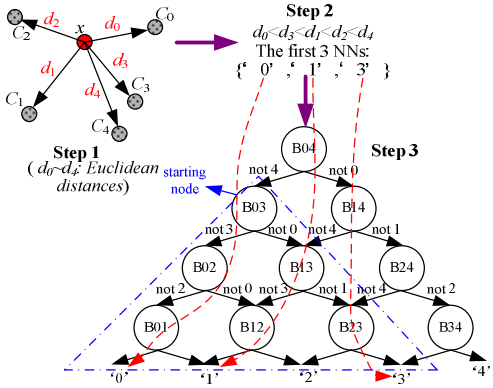

Fig. 3. An illustration of finding the starting node ($N = 5$, $K = 3$).

In this figure, three steps are implemented. Step 1 computes the Euclidean distances between $\boldsymbol{x}$ and the centroids, and five distances are obtained. These distances are sorted in the ascending order in Step 2 and the first 3 nearest neighbours (assumed to be '0', '1' and '3') are selected. In Step 3, we can observe that three selected classes, whose corresponding leaves are indicated in the final layer of DAG SVM in Fig. 3, are derived from the subsidiary DAG, enclosed in a triangle marked by dashed lines. The starting node (*i.e.* B03) is obtained as the root node of the sub graph.

Another way of obtaining the starting node is the use of the information included in indexes. In this case, the index for each class should be predefined and arranged according to the DAG SVM structure. For example, the index for class '0' is 0; for class '1' – 1; …; and so on. The indices i and j for a starting node Bij correspond to the minimal and maximal values of selected K numbers, respectively. Therefore, in a simple way, the starting node B03 can be obtained.

3.3. Classification system design based on SVM

The design of a classification system based on an SVM classifier is similar to that based on a neural network, and the steps of the presented *improved DAG SVM* (*i*DAG SVM) system design for PEC are as follows:

1) Feature extraction. The original current or voltage signals are sampled from the available sensors. These signals contain noise or redundant information, so they need to be pre-processed by signal processing techniques, such as PCA [15, 20, 31], FFT [2, 32], *wavelet transformation* (WT) [6, 7, 17], S-transformation [21], or Concordia transformation [9, 18]. This step generates feature samples, which can be used in the offline training.
2) Offline training of the SVM. Prior to the training, the feature samples need to be assigned with labels (+1 or −1). In our design, in the BSVM training of one-against-rest SVM, the feature samples corresponding to one class are labelled with −1 and the other faults' samples are labelled with +1. For the BSVM training of one-against-one SVM, features for the ith class are labelled with −1, and the features for the jth class are labelled with +1 ($i < j$).
   After training, for each BSVM, the generated SVM parameters are saved. Also, considering the *i*DAG SVM, the centroid of each class needs to be calculated and saved.
3) Fault classification. Given a new sample, the diagnostic flow is implemented according to Fig. 3.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

4. Case studies: rectifier and inverter

4.1. Simulated rectifier

The first circuit is a three-phase full-bridge rectifier with six uncontrolled diodes and a load resistor R_{load}, is shown in Fig. 4. This topology can be used in aerospace power systems and many industrial power electronic converter design applications.

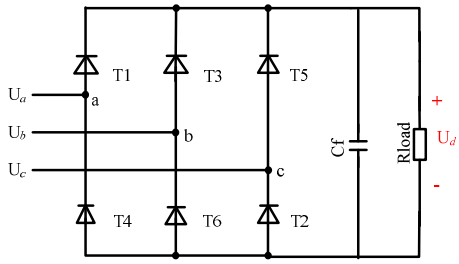

Fig. 4. The simulated three-phase full-bridge rectifier.

This circuit is modelled and simulated with Matlab R2010b-Simulink. In this simulation, the R_{load} value is set to 550 Ω with 10% tolerance (to simulate the load fluctuation) and the filter capacitor Cf value is set to 10 μF: the nominal phase frequency and voltage of the input source (U_a, U_b and U_c) are 400 Hz and 23 V rms, respectively. The output voltage U_d on the load is selected as an accessible signal.

In this circuit, open-circuit faults for the diodes are examined. Faulty diodes and their fault codes are listed in Table 1. $\{T_i, T_j\}$ ($i, j = 1,\ldots,6$ and $i \neq j$) means that two diodes T_i and T_j are faulty simultaneously. In this simulation, f0, indicating a sound circuit, is regarded as a special class. Hence, twenty two classes are considered. For each fault, this circuit model is simulated 50 times and each time the load value is varied and the corresponding signal U_d is sampled (a sample rate for the simulation is 20 kHz). In this way, altogether 50 samples for each fault can be collected. A randomly selected segment of sample for each fault is shown in Fig. 5.

Table 1. Faults for the rectifier.

Code	Faulty diode(s)	Code	Faulty diode(s)	Code	Faulty diode(s)
f0	–	f8	{T3, T6}	f16	{T1, T2}
f1	T1	f9	{T2, T5}	f17	{T2, T3}
f2	T2	f10	{T1, T3}	f18	{T3, T4}
f3	T3	f11	{T1, T5}	f19	{T4, T5}
f4	T4	f12	{T2, T4}	f20	{T5, T6}
f5	T5	f13	{T2, T6}	f21	{T6, T1}
f6	T6	f14	{T3, T5}		
f7	{T1, T4}	f15	{T4, T6}		

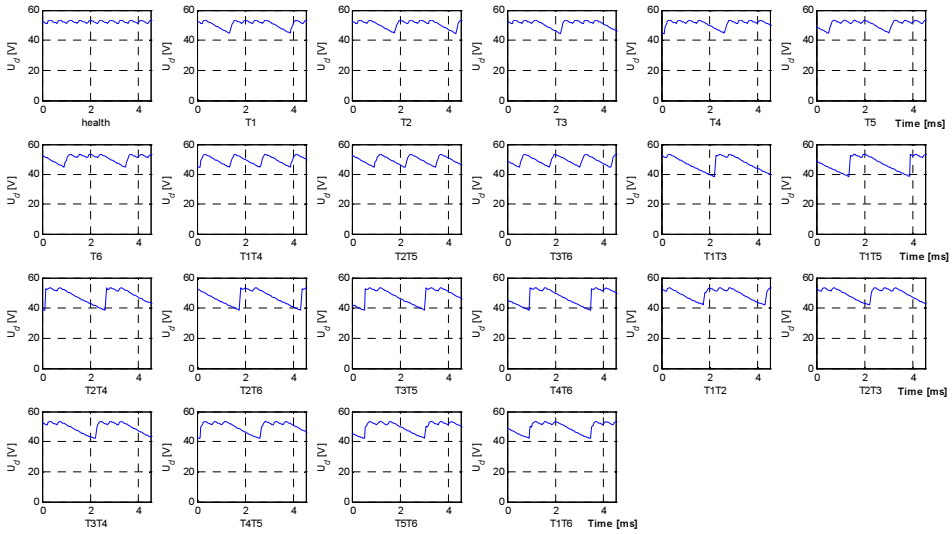

Fig. 5. Waveforms of faults for the simulated rectifier.

Feature extraction. In this example, the WT method is applied to U_d waveforms of each fault class to extract features. The WT is a useful technique [33] that can be used to decompose the collected data into the time-frequency domain, in which coarse coefficients and detail coefficients can be obtained. A simplified diagram of WT decomposition tree is shown in Fig. 6. The coarse coefficients in the low frequency band can indicate the outline of waveform. Generally, different waveforms, indicating different fault classes, can have different coarse coefficients. Hence, the coarse coefficients are selected as the fault features in our research. The detail coefficients in the high frequency band, however, are not considered as features, because these coefficients can be easily corrupted by noise.

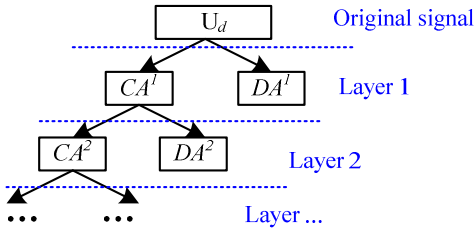

Fig. 6. A simplified diagram of WT decomposition tree.

The steps of extracting fault features with WT are as follows:

a) Apply WT to U_d, assumed to have N data points, and in each layer I (I = 1, 2, …) the coarse coefficients $CA^I = \{ CA^I(k) \}$ (k = 1, 2, …, $N/2I$) can be obtained after wavelet decomposition. Also, in WT applications, discussion of the number of decomposition layers and the mother function is inevitable. In our research, we determined these parameters by comparing different experiment results. Finally, *'Haar'* was selected as the mother function of WT for 5-layered decomposition, and these parameters can lead to good experiment results.

b) In each decomposition layer I, calculate the mean value of CA^I with the following equation:

$$Me^I = \frac{2^I}{N} \cdot \sum_{k=1}^{N/2^I} [CA^I(k)]. \tag{10}$$

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

Find the maximum value Mx^I from $(CA^I\text{-}Me^I)$, according to the following equation:

$$Mx^I = \max[\text{abs}\{CA^I(k) - Me^I\}]. \tag{11}$$

c) Normalize the coarse coefficients with 2-norm. In this case, normalization of coefficients can reduce the effect of load fluctuation. This step can be accomplished as follows:

$$\text{nCA}^I = \left\| \{ \frac{CA^I(k) - Me^I}{Mx^I} \} \right\|_2, \tag{12}$$
$$\left(k = 1,\ 2,\ \ldots,\ N/2I\ \right).$$

d) Calculate the feature nCA^I in layer I. In all, a fault feature vector $\mathbf{E}=[\text{nCA}^1, \text{nCA}^2, \text{nCA}^3, \text{nCA}^4, \text{nCA}^5]$ can be extracted. Finally, the feature vector **E** should be normalized to have zero mean and unity variance. In the machine learning domain, this is a commonly used means which can avoid a large data range.

Result. The Matlab codes for all classifiers were run on a P4 personal computer with 2.6 GHz dual CPUs and 2 GB RAM. Our operating system is Windows XP.

The feature set is split into two parts: a training set and a testing set. The training set contains 10 samples of each class, whereas 40 samples of each class are used for testing. A penalty parameter for SVM is set to 100, and σ is varied across a range {1, 2, 4, 8, 16, 32}. With these parameters, each SVM classifier can generate six results; the best result for each SVM is recorded. In this experiment, all SVM classifiers can achieve 100% testing accuracy when $\sigma = 1$. The *i*DAG SVM needs to confirm the value of *K*. In this research, we performed an exhaustive offline searching for *K*, ranging from 2 to 22, to find a suitable inflection point. In each search, the *K* value was changed, the *i*DAG SVM was used to evaluate the testing set, and accuracy was recorded. The testing accuracy, as well as the testing time, as functions of *K*, are shown in Figs. 7a–7b.

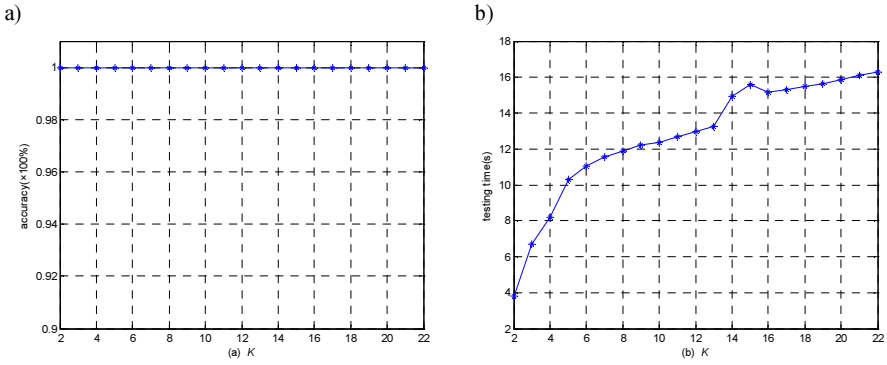

Fig. 7. Accuracy (a) and testing time (b), as functions of *K* for the simulated rectifier.

In Fig. 7a, we can observe that the accuracy curve is simple and any value of *K* can lead to 100% classification accuracy. Hence, we choose $K = 2$ as a suitable value, which gives the shortest testing time.

Five classifiers are compared in terms of the *classification Accuracy* (Acc), *training time* (TrT) and *testing time* (TeT) for the testing set. Comparison of SVM classifiers' performance is based on their best results ($\sigma = 1$) and shown in Table 2.

The BPNN used in this study is a forward-feed neural network with three-layered structure, whose training parameters are shown in Fig. 8. Training of this neural classifier was performed with Matlab toolbox for neural networks. Note that the number of hidden neurons is 26, with

which a good result can be obtained. Also, the neural classifier was trained for 3 times and the best result was added to Table 2. The Matlab function to validate the testing set is *'sim'*.

Table 2. Comparison of the classifiers' performance for the simulated rectifier.

Classifier	Acc	TrT [s]	TeT [s]
One-against-rest SVM	100%	3.2	19.1
One-against-one SVM	100%	5.2	174.3
DAG SVM	100%	5.2	15.6
*i*DAG SVM (*K* = 2)	100%	5.2	3.8
BPNN	100%	188.6	20.9

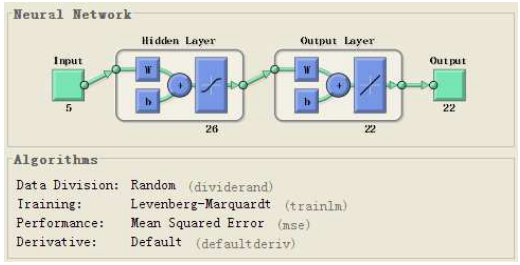

Fig. 8. The neural network used in our experiment.

4.2. Actual rectifier

An actual rectifier is mainly designed with six discrete power diodes (type: 6A10). A photo of the circuit fault diagnosis system (including a fault setup and a fault data acquisition) is shown in Fig. 9.

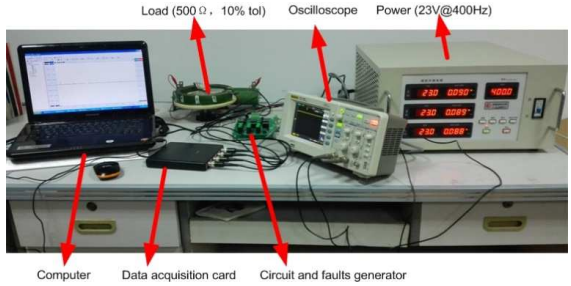

Fig. 9. A photo of the rectifier fault diagnosis platform.

In this system, each diode is connected in series with a relay and an action of the relay, manually controlled with a button, can generate an open fault of diode. A single fault can be generated by pushing one button, whereas double faults need pushing two buttons simultaneously. We used this system to collect 50 samples for each fault. Collecting a fault sample, each time we changed the load randomly within a 10% tolerance. The fault signals were collected with Handyscope HS4 (12-bit ADC inside), and a sample rate for this data collection device was set to 20 kHz, which was consistent with the setup during the simulation.

For each fault, a segment of one sample was randomly selected, as shown in Fig. 10. Subsequently, the WT was applied to these signals for feature extraction.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

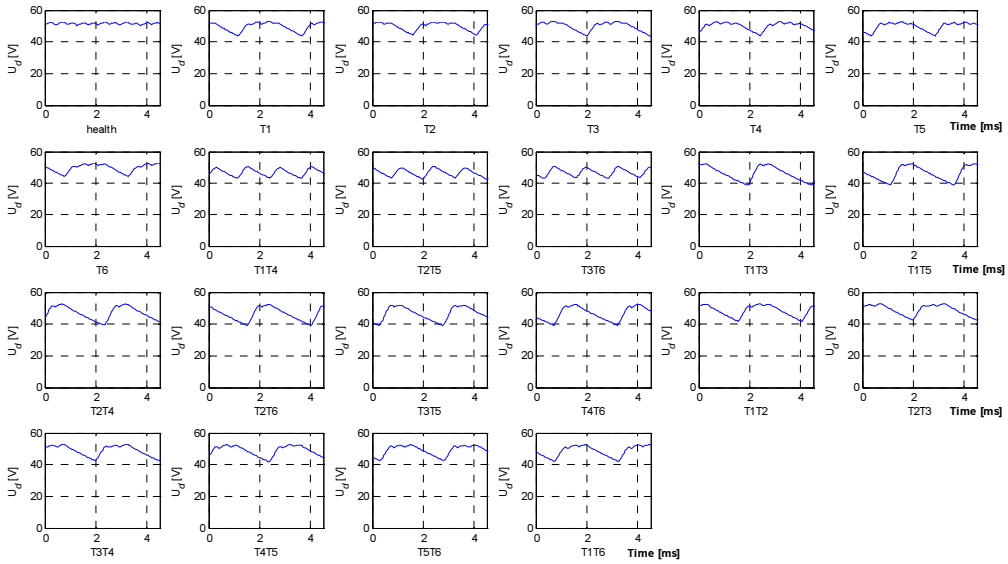

Fig. 10. Waveforms of faults for the actual rectifier.

The starting point (closely related to phase information) of the signal is important for WT, and we found the first starting point of the signal by zero-crossing detection of the input power source waveform U_{ab}. The basic principle of finding a starting point of U_d is shown in Fig. 11, which presents the waveforms of U_d and U_{ab} in a sound circuit condition.

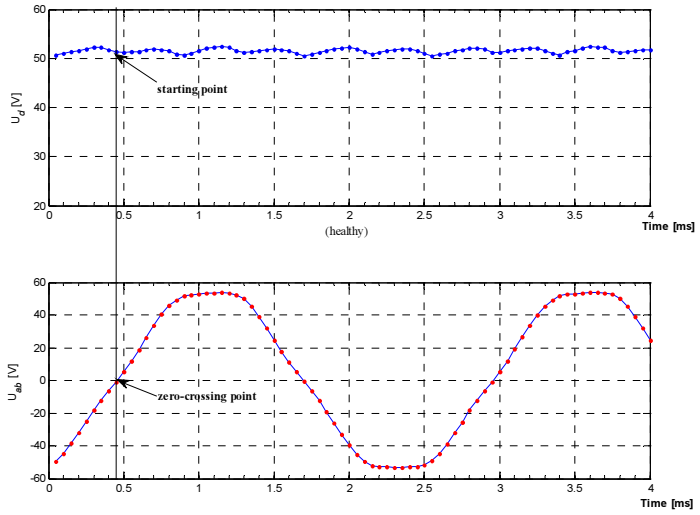

Fig. 11. Finding a starting point of U_d by using the zero-crossing detection method with U_{ab}.

In this research, the steps of feature extraction and classifiers' design were identical to those of the simulated rectifier. Fig. 12 presents the curves of accuracy and testing time for *i*DAG SVM. Comparison of the classifiers' performance with their best results is shown in Table 3 for $\sigma = 4$.

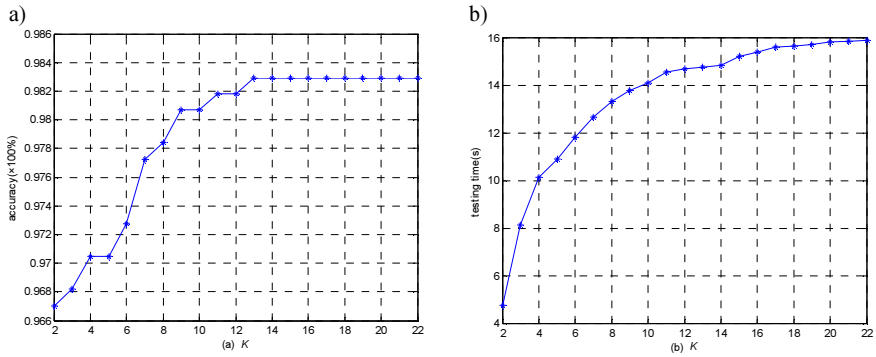

Fig. 12. Accuracy (a) and testing time (b), as functions of K for the actual rectifier with fluctuation of load (10%).

Table 3. Comparison of the classifiers' performance for the actual rectifier with fluctuation of load (10%).

Classifier	Acc	TrT [s]	TeT [s]
One-against-rest SVM	97.8%	2.6	18.6
One-against-one SVM	98.3%	5.1	184.3
DAG SVM	98.3%	5.1	17.0
*i*DAG SVM (K = 13)	98.3%	5.1	15.2
BPNN	96.8%	35.8	23.9

For the actual rectifier circuit we also performed another experiment, in which both the input power (including amplitude and frequency) and load were randomly fluctuated. The tolerances of input power amplitude, input frequency and load were set to 5%, 5% and 10%, respectively. This experiment aimed to examine the classifier performance in a complicated operation environment. The curves of finding a good result in the *i*DAG SVM classifier design are shown in Fig. 13 and, for $\sigma = 1$, several classifiers can achieve the best results, which are listed in Table 4.

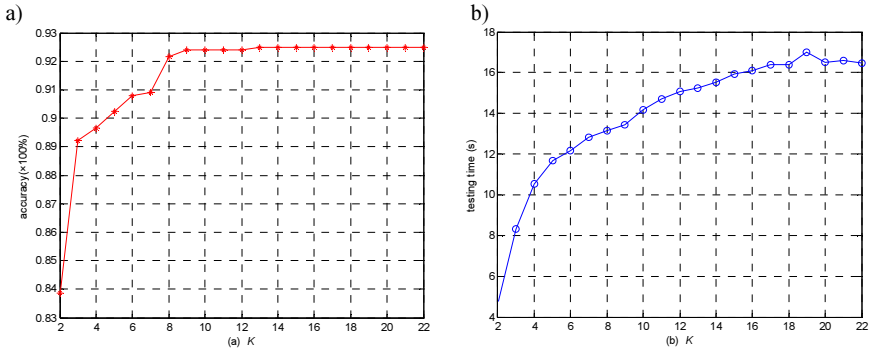

Fig. 13. Accuracy (a) and testing time (b), as functions of K for the actual rectifier with fluctuations of load (10%), input power amplitude (5%) and frequency (5%).

Table 4. Comparison of the classifiers' performance for the actual rectifier with fluctuations of load (10%), input power amplitude (5%) and frequency (5%).

Classifier	Acc	TrT [s]	TeT
One-against-rest SVM	91.1%	2.3	18.4
One-against-one SVM	93.2%	4.6	173.2
DAG SVM	92.5%	4.6	15.9
*i*DAG SVM (K = 12)	92.5%	4.6	14.9
BPNN	82.5%	13.1	20.7

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

4.3. Inverter

The third PEC, with its structure shown in Fig. 14, is an actual three-phase inverter which drives a motor with a symmetric structure. The motor is a 50 W *brushless DC motor* (BLDCM) with 3-phase and 5-pair poles, and the BLDCM shaft is coupled with an electric fan, running at a rate of 800 rpm (~5% fluctuation). This system is used for cooling in industry applications. Although the used inverter has a low driving power, the considered fault classification algorithms can be extended to inverters with a higher power in a straightforward way.

In this inverter, the MOSFETs are driven with a square wave *pulse width modulation* (PWM), and the voltage value is Vdc = 48 V. We consider a single switch open fault for this drive circuit, a total of 7 faults need to be classified. In this case, the fault code for switch T*i* is f*i* (*i* = 1, …, 6), and f0 denotes the sound condition of this circuit.

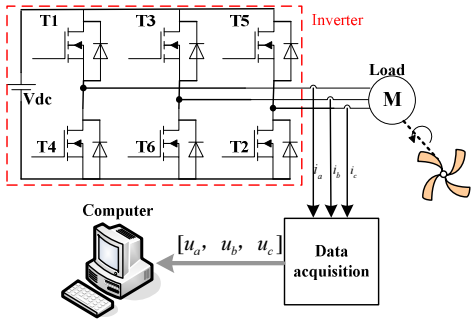

Fig. 14. The inverter system structure used in the experiment.

In this example, three phase currents (i_a, i_b and i_c) need to be collected synchronously. In this experiment, forty samples for each fault pattern were collected based on the experiment platform, in which an open fault of a power switch could be set by an emulator of the drive circuit controller.

Feature extraction. The presented feature extraction algorithm needs two steps.

Step 1: The WT is adopted to reduce the effect of noise. *'Haar'* wavelet is selected as the mother function to decompose the currents' signals into coarse coefficients ($i_a^w(k), i_b^w(k), i_c^w(k)$) and detail coefficients in layer 3. The detail coefficients were discarded in this design.

In order to reduce the effect of load, with reference to [17], normalization of wavelet coefficients can be considered:

$$\begin{aligned} \hat{i}_a(k) &= i_a^w(k) / \max(\text{abs}(i_a^w)) \\ \hat{i}_b(k) &= i_b^w(k) / \max(\text{abs}(i_b^w)) \\ \hat{i}_c(k) &= i_c^w(k) / \max(\text{abs}(i_c^w)) \\ k &= 1, 2, \ldots, M/8, \end{aligned} \tag{13}$$

where M is the number of collected data points.

The segments of collected phase currents for each fault before and after *'Haar'* WT in layer 3 are shown in Figs. 15a–15g, respectively.

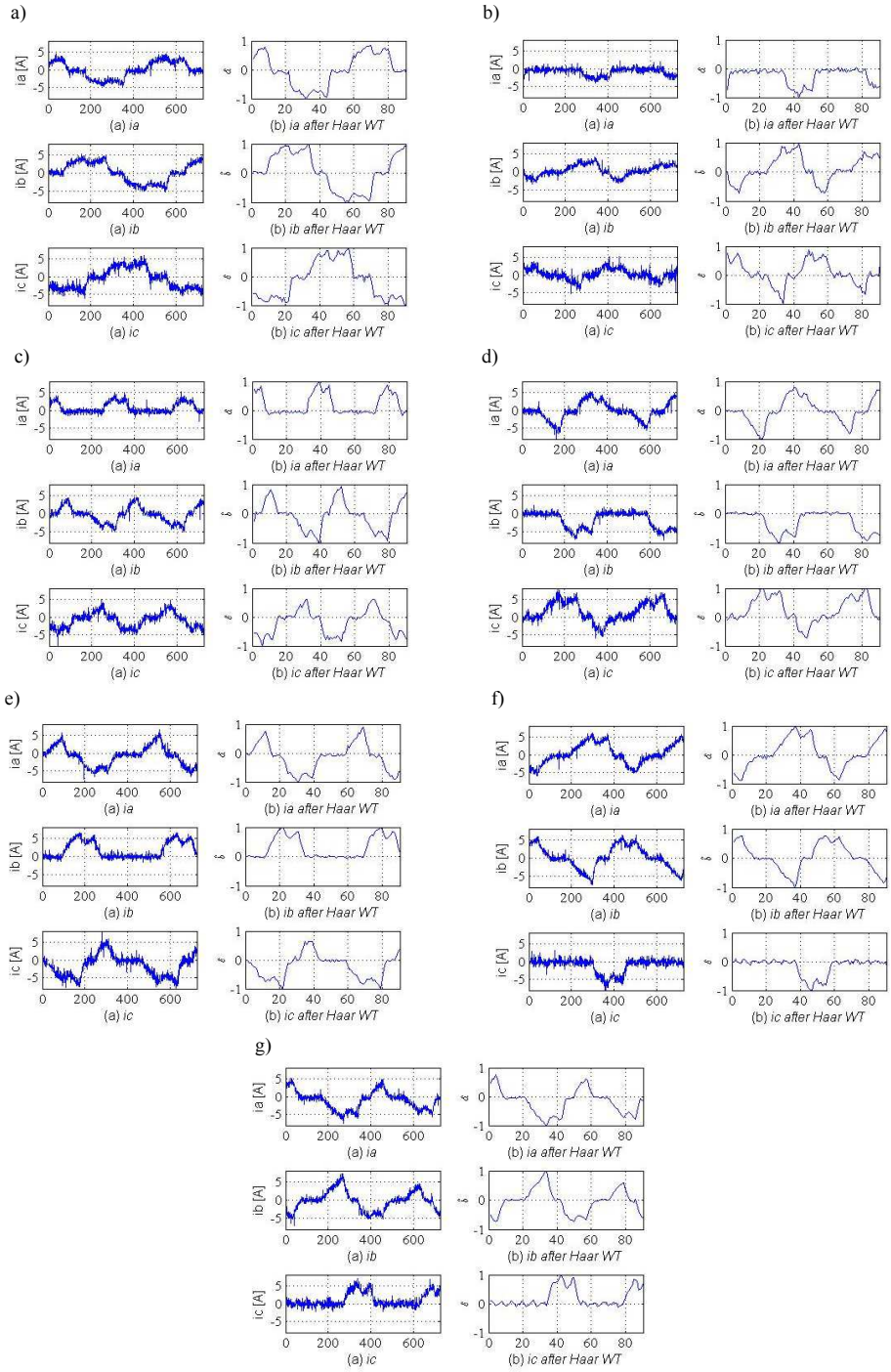

Fig. 15. Current waveforms before and after adapting WT for seven faults, respectively. Waveforms under no fault (a); waveforms under T1 fault (b); waveforms under T2 fault (c); waveforms under T3 fault (d); waveforms under T4 fault (e); waveforms under T5 fault (f); waveforms under T6 fault (g).

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

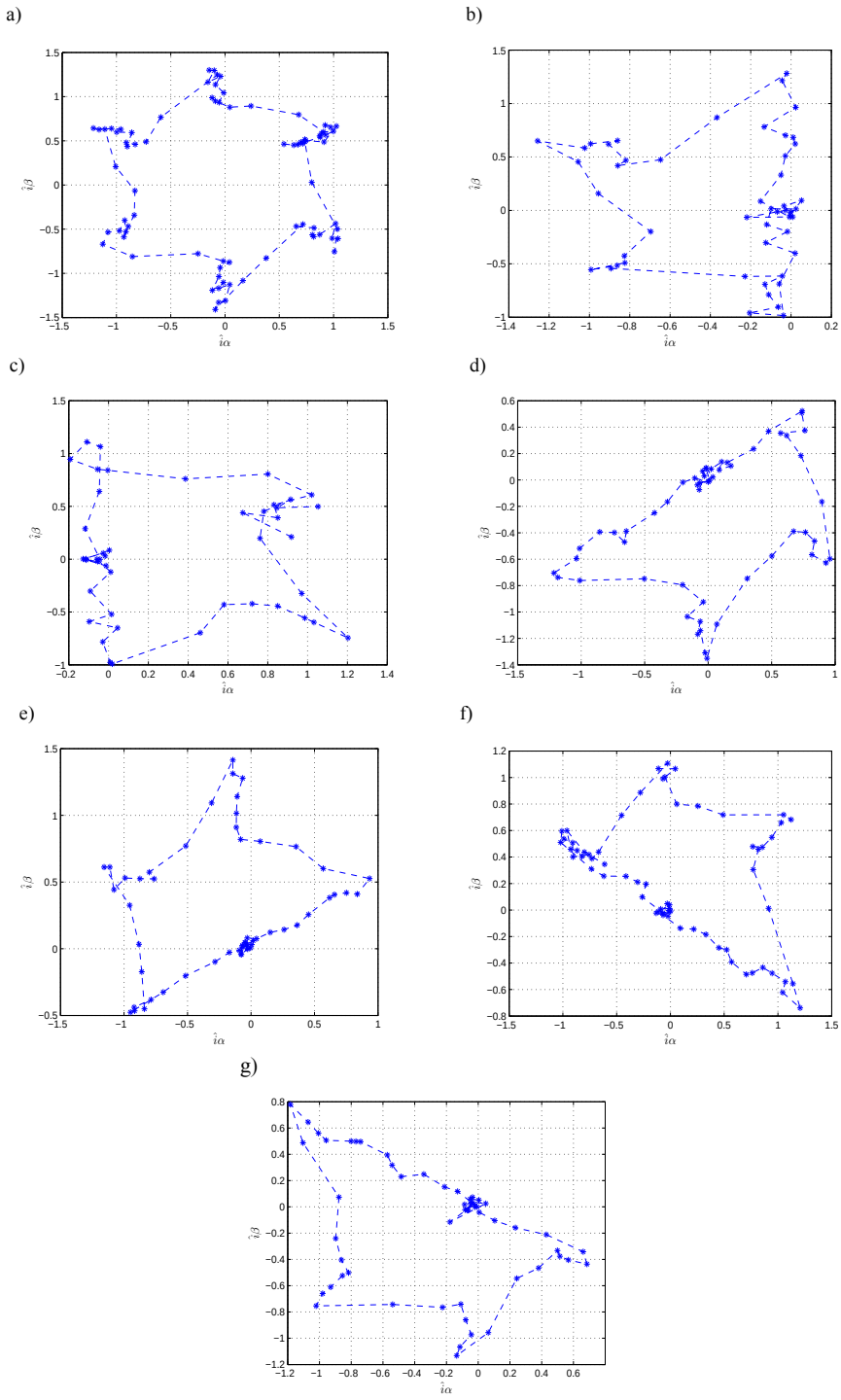

Fig. 16. Current trajectories for seven faults, respectively. A current trajectory under no fault (a); a current trajectory under T1 fault (b); a current trajectory under T2 fault (c); a current trajectoryunder T3 fault (d); a current trajectory under T4 fault (e); a current trajectory under T5 fault (f); a current trajectory under T6 fault (g).

In the Fig. 15, we can observe that, after decomposition of 3 layers, the outlines of current waveforms are clear. Also, we can find that the waveforms for each fault are different from others, and hence they can be used for subsequent fault classification.

Step 2: For the AC motor system with a symmetric structure, the Concordia transform can be used to calculate a 2-dimensional current trajectory ($\hat{i}_\alpha, \hat{i}_\beta$) in the α-β frame:

$$\begin{cases} \hat{i}_\alpha = (2\hat{i}_a - \hat{i}_b - \hat{i}_c)/\sqrt{6} \\ \hat{i}_\beta = (\hat{i}_b - \hat{i}_c)/\sqrt{2} \end{cases}. \tag{14}$$

Figure 16 illustrates selected trajectories for ($\hat{i}_\alpha, \hat{i}_\beta$) under faults. Note that the numbers on the axes have no units because they are ratios from the formulae (13) and (14). We can observe that these trajectories are different from each other mainly in terms of geometry. Hence, we can consider extracting simple centroid features from these trajectories.

The centroid features can be extracted from a closed trajectory [34, 35]:

$$\begin{cases} C_\alpha = \dfrac{8}{M}\displaystyle\sum_{k=1}^{M/8} \hat{i}_\alpha(k) \\ C_\beta = \dfrac{8}{M}\displaystyle\sum_{k=1}^{M/8} \hat{i}_\beta(k) \\ r = \dfrac{8}{M}\displaystyle\sum_{k=1}^{M/8} \sqrt{(\hat{i}_\alpha(k) - C_\alpha)^2 + (\hat{i}_\beta(k) - C_\beta)^2} \\ \hat{C}_\alpha = C_\alpha / r \\ \hat{C}_\beta = C_\beta / r \end{cases} \tag{15}$$

In order to obtain a completely closed trajectory, M should be the number of data points from at least one cycle of waveform. A 2-dimensional feature vector $\mathbf{E}=[\hat{C}_\alpha, \hat{C}_\beta]$ can describe the centroid of trajectory regarding its radius size. The 2-dimensional centroid features are sufficient to discern seven faults and can lead to good classification results.

Results for inverter. Fourteen feature samples were used as the training set, whereas the other 26 samples were used as the validation set. By confining C to 100, the one-against-rest SVM can achieve the best result for $\sigma=1$, whereas the one-against-one SVM, – for $\sigma=2$.

For the designed *i*DAG SVM, the validation set is also used to search for a proper value of K. The searching curves are given in Fig. 17.

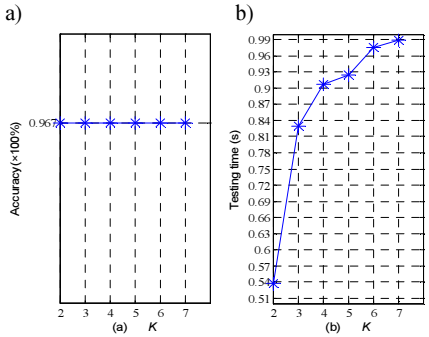

Fig. 17. Accuracy (a); testing time (b), as functions of K.

In this experiment, different values of K can lead to the same accuracy (96.7%). Hence, K = 2 was directly selected for this *i*DAG SVM. In the neural classifier design, the number of hidden neurons is set to 7 and the BPNN is trained for several times. The best result of 97.8% is recorded. Good results for these classifiers are shown in Table 5.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 701–720.

Table 5. Comparison of the classifiers' performance for the inverter.

Classifier	Acc	TrT [s]	TeT [s]
One-against-rest SVM	96.1%	1.0	1.12
One-against-one SVM	96.7%	0.6	3.66
DAG SVM	96.7%	0.6	0.99
*i*DAG SVM (*K* = 2)	96.7%	0.6	0.54
BPNN	97.8%	0.3	4.65

4.4. Analysis

1) According to the above examples, we believe that the SVM and neural classifiers are applicable to power electronic circuit faults' classification, if only the parameter values are properly chosen.
2) Through several experiments, we found that the neural networks can achieve good results when classifying a small number of faults. However, it is not a good choice to classify a large number of faults with conventional neural networks, because these classifiers will need much more hidden neurons to implement a complex training process and so to increase the calculation complexity. Changing a big network into some small-scale networks can solve this problem [32], but both the structure and parameters of these small networks need to be determined deliberately. Moreover, the neural network classifier exhibits different performance for different trainings.
   The SVM classifiers can be regarded as alternative solutions for the neural classifiers in the applications to power electronic system diagnosis, because an SVM classifier can achieve very close or even better performance to that of a neural classifier. In addition, a standard SVM classifier can exhibit stable performance for different trainings. Moreover, it needs to tune relatively fewer parameters.
3) The SVM classifier has many forms [36, 37]. In the research, we examined two typical forms in the domain of power electronic circuit fault diagnosis. As a result, the one-against-one SVM is usually used for classification, because it can achieve a high classification accuracy. However, this classifier needs a high computational complexity to implement classification and this drawback limits its usage in some applications, *e.g.* in power electronic circuit on-line fault diagnosis, surveillance or even fault-tolerant systems. Hence, the structure of this classifier needs to be improved or rearranged to adapt to fast fault classification. We adopt the DAG SVM as an alternative for this classifier.
   In our research, the DAG SVM and the one-against-one SVM were compared in terms of classification accuracy and testing time, and we found that the DAG SVM's performance was very close to its counterpart's, but with a significantly lower computational complexity needed. Hence, we believe that the DAG SVM can be used to replace the one-against-one SVM for fault classification of PECs. The *i*DAG SVM, with the help of nearest neighbours, can further reduce the testing time, whilst maintaining almost unchanged performance.
4) The WT is a good tool for fault feature extraction of power electronic circuits. We achieved good results by using this tool in the experiments. The proper selection of a good mother function and decomposition layers is a difficult problem, and in our research we solved this problem by comparing and evaluating the experiment results with different parameters.

5. Conclusions

In industrial applications, a fast method of fault diagnosis in power electronic circuits is important because of the requirements of high reliability and fault-tolerance. This paper presents a data-driven method of fault classification in power electronic circuits, and this method is based on the DAG SVM. This classifier can be improved by combining it with the

K-NN method. Compared with the conventional one-against-rest SVM and one-against-one SVM, the presented method has a very high implementation speed, because this method is based on the DAG SVM, which needs to compute (N–1) BSVMs for N faults. After the improvement, the number of BSVMs needed by *i*DAG SVM is less than or equal to (N–2). Hence, among the SVM classifiers, the *i*DAG SVM has the lowest computational complexity. Also, the *i*DAG SVM has the classification performance comparable with that of the DAG SVM, if only the parameter K is properly selected. In our research, K was determined on the basis of experimental searching results. In another way, K can be selected with an empirical formula as follows:

$$K = \lceil N/2 \rceil. \tag{16}$$

With this formula, in many experiments, we found that the *i*DAG SVM can achieve satisfactory results. Note that the proposed classifier can also be considered as a general method, and this classifier can be easily extended to other fast fault classification applications. The presented method also has some limitations, because it is based on the conventional DAG SVM, which needs to be prearranged. In arranging the DAG SVM structure, the selection of a root node for the DAG SVM is a problem. Different root nodes will probably lead to different performance of the classifier. Hence, the proper selection of a root node should be further studied. In the future, we can consider some available methods in designing pattern classifiers to solve this problem [38–40].

Finally, the feature extraction is important in the design of a successive classifier In our research we need to select fault features manually, so in the future work, automatic and efficient feature selection methods will be examined.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant # 51377079).

References

[1] Mohagheghi, S., Harley, R.G., Habetler, T.G., Divan, D. (2009). Condition monitoring of power electronic circuits using artificial neural networks. *IEEE Trans. Power Electron.*, 24(10), 2363–2367.

[2] Khomfoi, S., Tolbert, L.M. (2007). Fault diagnostic system for a multilevel inverter using a neural network. *IEEE Trans. Power Electron.*, 22(03), 1062–1069.

[3] Mirafzal, B. (2014). Survey of fault-tolerance techniques for three-phase voltage source inverters. *IEEE Trans. Ind. Electron.*, 61(10), 5192–5202.

[4] Filippetti, F., Franceschini, G., Tassoni, C., Vas, P. (2000). Recent developments of induction motor drives fault diagnosis using AI techniques. *IEEE Trans. Ind. Electron.*, 47(05), 994–1004.

[5] Zidani, F., Diallo, D., Benbouzid, M.E.H., Naït-Saïd, R. (2008). A fuzzy-based approach for the diagnosis of fault modes in a voltage-fed PWM inverter induction motor drive. *IEEE Trans. Ind. Electron.*, 55(02), 586–593.

[6] Khanniche, M.S., Mamat-Ibrahim, M.R. (2004). Wavelet-fuzzy-based algorithm for condition monitoring of voltage source inverter. *Electron. Lett.*, 40(04), 267–268.

[7] Potamianos, P.G., Mitronikas, E.D., Safacas, A.N. (2014). Open-circuit fault diagnosis for matrix converter drives and remedial operation using carrier-based modulation methods. *IEEE Trans. Ind. Electron.*, 61(1), 531–545.

[8] An, Q.T., Sun, L.Z., Sun, L., Jahns, T.M. (2010). Low-cost diagnostic method for open-switch faults in inverters. *Electron. Lett.*, 46(14), 1021–1022.

[9] Diallo, D., Benbouzid, M.E.H., Hamad, D., Pierre, X. (2005). Fault detection and diagnosis in an induction machine drive: a pattern recognition approach based on Concordia stator mean current vector. *IEEE Trans. Energy Conver.*, 20(03), 512–519.

[10] Charfi, F., Sellami, F., Al-Haddad, K. (2006). Fault diagnostic in power system using wavelet transforms and neural networks. *Proc. ISIE*, 1143–1148.

[11] Kadri, F., Drid, S., Djeffal, F.Y., Chrifi-Alaoui, L. (2013). Neural classification method in fault detection and diagnosis for voltage source inverter in variable speed drive with induction motor. *Proc. EVER*, 1–5.

[12] Masrur, M.A., Chen, Z., Murphey, Y. (2010). Intelligent diagnosis of open and short circuit faults in electric drive inverters for real-time applications. *IET Power Electron.*, 3(02), 279–291.

[13] Ma, C., Gu, X., Wang, Y. (2009). Fault diagnosis of power electronic system based on fault gradation and neural network group. *Neurocomputing*, 72(13–15), 2909–2914.

[14] Lu, B., Sharma, S.K. (2009). A literature review of IGBT fault diagnostic and protection methods for power inverters. *IEEE Trans. Ind. Appl.*, 45(5), 1770–777.

[15] Khomfoi, S., Tolbert, L.M. (2007). Fault diagnosis and reconfiguration for multilevel inverter drive using AI-based techniques. *IEEE Trans. Ind. Electron.*, 54(6), 2954–2968.

[16] Fan, B., Dong, M., Zhao, J., Zhang, Q. (2010). Three-phase inverter fault diagnosis based on optimized neural networks. *Proc. ICCASM*, 4, 482–485.

[17] Kim, D.E., Lee, D.C. (2008). Fault diagnosis of three-phase PWM inverters using wavelet and SVM. *Proc ISIE*, 329–334.

[18] Delpha, C., Chen, H., Diallo, D. (2012). SVM based diagnosis of inverter fed induction machine drive: a new challenge. *Proc. IECON*, 3931–3936.

[19] Hsu, C.W., Lin, C.J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Networ.*, 13(2), 415–425.

[20] Wang, R., Zhan, Y., Zhou, H., Cui, B. (2013). A fault diagnosis method for three-phase rectifiers. *Int. J. Elec. Power*, 52, 266–269.

[21] Wang, R., Zhan, Y., Zhou, H. (2012). Application of S transform in fault diagnosis of power electronics circuits. Scientia Iranica, 19(3), 721–726.

[22] Xu, H., Zhang, J., Qi, J., Wang, T., Han, J. (2014). RPCA-SVM fault diagnosis strategy of cascaded H-bridge multilevel inverters. *Proc. ICGE*, 164–169.

[23] Hu, Z., Gui, W., Yang, C., Deng, P., Ding, S. X. (2011). Fault classification method for inverter based on hybrid support vector machines and wavelet analysis. *Int. J. Control Autom. Syst.*, 9(4), 797–804.

[24] Huang, C., Zhao, J., Wu, C. (2013). Data-based inverter IGBT open-circuit fault diagnosis in vector control induction motor drives. *Proc. IEEE ICIEA*, 1039–1044.

[25] Cortes, C., Vapnik, V.N. (1995). Support-vector networks. *J. Mach. Learn.*, 20(3), 273–297.

[26] Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Trans. Neural Networ.*, 10(5), 988–999.

[27] Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.*, 2(2), 121–167.

[28] Chapelle, O., Haffner, P., Vapnik, V.N. (1999). Support Vector Machines for histogram-based image classification. *IEEE Trans. Neural Networ.*, 10(5), 1055–1064.

[29] Platt, J.C., Cristianini, N., Shawe Taylor, J. (2000). Large margin DAGs for multi-class Classification. *Advances in Neural Information Processing Systems*, 12, 547–553.

[30] Biet, M. (2013). Rotor faults diagnosis using feature selection and nearest neighbors rule: application to a turbogenerator. *IEEE Trans. Ind. Electron.*, 60(9), 4063–4073.

[31] Martins, J.F., Pires, V.F., Lima, C., Pires A.J. (2012). Fault detection and diagnosis of grid-connected power inverters using PCA and current mean value. *Proc. IECON*, 5185–5190.

[32] Murphey, Y.L., Masrur, M.A., Chen, Z., Zha, B. (2006). Model-based fault diagnosis in electric drives using machine learning. *IEEE-ASME Trans. Mech.*, 11(3), 290–303.

[33] Mallat, S.G. (1989). A theory for multi-resolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal*., 11(7), 674–693.

[34] Fernao, V.P., Amaral, T.G., Martins, J.F. (2012). Fault detection and diagnosis of voltage source inverter using the 3D current trajectory mass center. *Proc. IEEE ICIT*, 737–742.

[35] Cui, J. (2015). Faults classification of power electronic circuits based on a support vector data description method. *Metrol. Meas. Syst*., 22(2), 205–222.

[36] Cui, J., Wang, Y. (2011). A novel approach of analog circuit fault diagnosis using support vector machines classifier. *Measurement*, 44(1), 281–289.

[37] Cui, J., Wang, Y. (2011). Analog circuit fault classification using improved one-against-one support vector machines. *Metrol. Meas. Syst*., 18(4), 569–582.

[38] Gu, B., Sheng, Victor S., Li, S. (2015). Bi-parameter space partition for cost-sensitive SVM. *Proc. IJCAI*, 3532–3539.

[39] Wen, X., Shao, L., Xue, Y., Fang, W. (2015). A rapid learning algorithm for vehicle classification. *Inform. Sciences*, 295(1), 395–406.

[40] Gu, B., Sheng, Victor S. (2016). A Robust Regularization Path Algorithm for ν-Support Vector Classification. *IEEE Trans. Neural Netw. Learn. Syst*., DOI: 10.1109/TNNLS.2016.2527796.

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 721–728.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

FAST SECOND ORDER ORIGINAL PRONY'S METHOD FOR EMBEDDED MEASURING SYSTEMS

Jarosław Zygarlicki

Opole University of Technology, Faculty of Electrical Engineering, Automatic Control and Informatics, Prószkowska 76, 45-758 Opole, Poland (✉ j.zygarlicki@po.opole.pl, +48 77 449 8074)

Abstract

The paper presents a method of adaptation of the original second order Prony's method for applications in low-cost digital measurement systems with low computing performance. The presented method can be used in measuring systems where it is important to obtain in real time the values of amplitude, frequency, initial phase and damping coefficient of a single sinusoidal component of an analysed signal. The paper presents optimized, in terms of the number of mathematical operations, implementation of the method in selected embedded devices as well as the calculation times of the method for each platform.

Keywords: Prony's method, signal processing, harmonics, measurements.

1. Introduction

Analysis of a digital multi-frequency signal to estimate the basic parameters of its components is a very broad subject mainly related to Fourier transforms [1–7]. One section of this area includes the analysis dedicated to estimation of a single component with the greatest possible accuracy and short calculation time. This area of application includes some modifications of the Fourier analysis [8–9] and other methods [10], of which the Prony's methods gain greater and greater practical significance [11–16].

The Prony's methods are characterized by the properties of precise estimation of the parameters of an analysed signal. They generate new measurement possibilities by identifying real frequencies of the analysed signal components, and by extending the signal model with information about the damping coefficients of the components [17–20]. These methods enable the use of short windows of analysis, which is valuable in the study of fast variable phenomena. They also specify the parameters of components of slowly variable signals, with incomplete periods in the analysed analysis window. Nevertheless, when analysing multiple components, they are computationally complex methods that require inversion of large matrixes, and calculation of roots of high-order polynomials. These methods may also involve problems with the numerical stability of solutions.

The great versatility of Prony's methods makes them an alternative to Fourier transform-based methods, enabling to measure a wider range of signals in variable measurement conditions (analysis window duration, sampling frequency) not available with other methods.

The paper presents a method of modifying the calculations required in the algorithm of original version of Prony's method of the second order in such a way as to obtain maximum simplification. The proposed modification is based on fundamental mathematical operations without involving complex numbers and operations requiring iterative calculations. This enables a significant reduction of the calculation time of Prony's method even for low-performance embedded devices. In the paper the accuracy aspect of calculations of Prony's

Article history: received on May 11, 2017; accepted on Jun. 24, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0058.

method is deliberately omitted, as it is the subject of separate publications [17–20]. The implementation of the method for a specific application is also not presented so as not to narrow down the group of potential recipients of the proposed solution.

The paper consists of 4 Sections. In Section 1 an introduction is included. In Sector 2 there are described the original Prony's method and its modifications to simplify the calculations of the presented algorithm. Section 3 shows the implementation of the method in selected embedded devices and the measurement results of algorithm execution time. Sector 4 contains a summary.

2. Description of adaptation of original Prony's method for embedded applications

The original Prony's method can be presented essentially as two calculation stages. In the first stage, frequency and damping coefficients of the complex exponents modelling the analysed signal are determined. In the second stage, amplitudes and initial phases of the components are calculated based on the parameter values determined in the previous stage [17, 21].

2.1. Determination of frequency and damping coefficients of components

In the first stage of the original Prony's method calculations a Toeplitz matrix is created on the basis of samples $x_1 \ldots x_{2p}$ of analysed signal, (left side of (1)). The following equation is based on this matrix:

$$\begin{bmatrix} x_p & x_{p-1} & \cdots & x_1 \\ x_{p+1} & x_p & \cdots & x_2 \\ \vdots & \vdots & \cdots & \vdots \\ x_{2p-1} & x_{2p-2} & \cdots & x_p \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_p \end{bmatrix} = - \begin{bmatrix} x_{p+1} \\ x_{p+2} \\ \vdots \\ x_{2p} \end{bmatrix}, \tag{1}$$

where p is a size of Prony's model, and the vector $A_1 \ldots A_p$ is a set of certain coefficients, which will be described later in the argument. For the adopted order of Prony's model $p = 2$, the solution of (1) can be represented by a relation:

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = - \left(\begin{bmatrix} x_p & x_{p+1} \\ x_{p-1} & x_p \end{bmatrix} \cdot \begin{bmatrix} x_p & x_{p-1} \\ x_{p+1} & x_p \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} x_p & x_{p+1} \\ x_{p-1} & x_p \end{bmatrix} \cdot \begin{bmatrix} x_{p+1} \\ x_{p+2} \end{bmatrix}. \tag{2}$$

To increase the transparency, let, $x_1 = a$, $x_2 = b$, $x_3 = c$, $x_4 = d$. By making further transformations related to the calculation of the inverse matrix we obtain:

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = -\frac{1}{r} \begin{bmatrix} a^2 + b^2 & -ab - bc \\ -ab - bc & b^2 + c^2 \end{bmatrix} \cdot \begin{bmatrix} b & c \\ a & b \end{bmatrix} \cdot \begin{bmatrix} c \\ d \end{bmatrix}, \tag{3}$$

where:

$$r = \left(a^2 + b^2\right)\left(b^2 + c^2\right) - \left(ab + bc\right)^2 \tag{4}$$

is a determinant of the inverted matrix. Estimation of component parameters of Prony's model takes place for $r \neq 0$. Finally, it can be written as:

$$A_1 = -\frac{bc\left(a^2 + b^2\right) - ac\left(ab + bc\right) + cd\left(a^2 + b^2\right) - bd\left(ab + bc\right)}{r}, \tag{5}$$

$$A_2 = -\frac{ac\left(b^2 + c^2\right) - bc\left(ab + bc\right) + bd\left(b^2 + c^2\right) - cd\left(ab + bc\right)}{r}. \tag{6}$$

In this way coefficients of a polynomial of the general form are determined:

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 721–728.

$$\phi(z)=\sum_{m=0}^{p} A_m z^{p-m} \tag{7}$$

for which the next step is to determine its complex roots z_k:

$$\phi(z)=\prod_{k=1}^{p}(z-z_k). \tag{8}$$

It is assumed that A_0 =1 [21]. For the considered case p = 2, a square polynomial of the general form is obtained:

$$\phi(z)=A_0 z^2+A_1 z+A_2, \tag{9}$$

where, knowing the coefficients A_0, A_1, A_2, the zero of the function can be determined using the commonly known Vieta's formulas. For the polynomial (9) we can first write:

$$\Delta = A_1^{\,2}-4A_0A_2. \tag{10}$$

The estimation of the sinusoidal components of Prony's model is obtained for complex roots, *i.e.* for $\Delta<0$. For the Prony's model with p = 2, we obtain conjugate roots describing sinusoidal components damped: one with a positive frequency and the other with identical amplitude, initial phase and damping but with a negative frequency. By further transformation of Vieta's formulas, we obtain the solution:

$$\mathrm{Re}\{z_1\}=\mathrm{Re}\{z_2\}=-\frac{A_1}{2A_0}, \tag{11}$$

$$\mathrm{Im}\{z_1\}=-\mathrm{Im}\{z_2\}=\frac{\sqrt{-\Delta}}{2A_0}. \tag{12}$$

The roots in the first stage of Prony's method are complex but for further embedded applications the operations can be performed for real data types, as – based on the real part $\mathrm{Re}\{z_1\}$ and imaginary part $\mathrm{Im}\{z_1\}$ of selected individual root – the frequencies f_1 and f_2 of the components can be calculated according to the following relation:

$$|z_1|=|z_2|=\sqrt{\frac{A_1^{\,2}-\Delta}{4A_0^{\,2}}}, \tag{13}$$

$$f_1=-f_2=\frac{1}{2\pi T}\arcsin\left(\frac{\mathrm{Im}\{z_1\}}{|z_1|}\right), \tag{14}$$

where T is a sampling period of the analysed signal, whereby, if the estimated component is not damped, then $|z_1|=|z_2|=1$. The damping coefficients α_1 and α_2 can be determined from the relation:

$$\alpha_1=\alpha_2=\frac{1}{T}\ln(|z_1|). \tag{15}$$

Finally, by making simple transformations, we can determine the frequency and damping coefficients of an estimated component using a simple C code or Matlab:

```
a2=a*a; b2=b*b; c2=c*c; ab=a*b;
bc=b*c; cd=c*d; ac=a*c; bd=b*d;
a2_b2=a2+b2; b2_c2=b2+c2; ab_bc=ab+bc;
r=a2_b2*b2_c2-ab_bc*ab_bc;
A1=-(bc*a2_b2-ac*ab_bc+cd*a2_b2-bd*ab_bc)/r;
A2=-(ac*b2_c2-bc*ab_bc+bd*b2_c2-cd*ab_bc)/r;
del=A1*A1-4*A2; rez1=-A1/2; imz1=sqrt(-del)/2;
absz1=sqrt(rez1*rez1+imz1*imz1);
F=asin(imz1/absz1)/(2*pi*T);
AL=log(absz1)/T;
```

For calculation of the first stage of Prony's model $p = 2$ in the original version, it is required to perform in total: 24 multiplications, 12 additions, 5 divisions, and additionally 2 root extractions and 1 arcsin operation. All operations are performed on real data types.

2.2. Determination of initial stages and amplitudes

In the second stage of the original Prony's method, the first operation is to determine a Vandermonde matrix (16):

$$V = \begin{bmatrix} z_1^0 & z_2^0 & \cdots & z_p^0 \\ z_1^1 & z_2^1 & \cdots & z_p^1 \\ \vdots & \vdots & \cdots & \vdots \\ z_1^{p-1} & z_2^{p-1} & \cdots & z_p^{p-1} \end{bmatrix} \tag{16}$$

from complex roots z_k and solving the (17):

$$h = \left(V^{\mathrm{T}}V\right)^{-1}V^{\mathrm{T}}x, \tag{17}$$

where: $h = [h_1,\ldots,h_p]^{\mathrm{T}}$, and $x=[x_1,\ldots,x_p]^{\mathrm{T}}$.

The determined vector h is used in the next step to calculate amplitudes *amp* and initial stages φ of the components of Prony's model, according to the relation [21]:

$$amp = |h|, \tag{18}$$

$$\varphi = \arcsin\left(\frac{\mathrm{Im}\{h\}}{|h|}\right). \tag{19}$$

When using a model with size $p = 2$, the number of mathematical operations to perform is relatively small. However, multiplications and additions on complex numbers are required, which greatly increases the number of operations based on real data types. For example, 1 complex multiplication translates into 4 multiplications and 2 additions of real numbers, and one complex addition translates into 2 additions of real numbers. Therefore, the direct adaptation of the second stage of Prony's method for embedded applications is not favourable in terms of the number of mathematical operations.

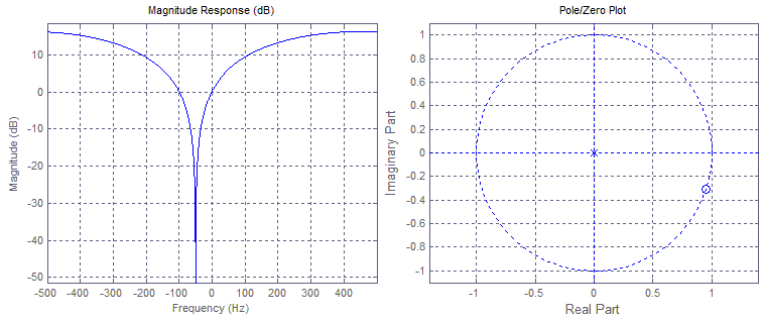

Fig. 1. The gain characteristics and positions of zeros of an FIR filter with coefficients C_{row1} calculated for a sinusoidal signal $f = 50$ Hz, $\alpha = 0$ with $T = 1$ ms.

The key to reducing the computational complexity of this stage lies in the observation that the term $(V^{\mathrm{T}}V)^{-1}V^{\mathrm{T}}$ in the relation (17) is a certain constant for the frequencies and coefficients of damping with a p x p matrix (C). The last step in calculation of the relation (17) is therefore

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 721–728.

the product of the matrix C and the vector of analysed signal x. For $p = 2$, this operation can be replaced by two operations of vector multiplication:

$$h_1 = C_{row1}x \text{ and } h_2 = C_{row2}x, \tag{20}$$

whereby, the solution $h_1 = \text{conj}(h_2)$ is obtained in the same way as $z_1 = \text{conj}(z_2)$. Therefore, it is enough to determine the parameters of a single element of vector h. Note that the operations from (20) are filter operations with FIR type filters with coefficients: C_{row1} for h_1 of signal x and C_{row2} for h_2 of signal x. However, the complex calculation of single FIR filter coefficients is still required, as described above. Further observations can be made by observing the performance characteristics of FIR filter with coefficients C_{row1} – Fig. 1.

It turns out that for the original Prony's method with $p = 2$, if $\alpha = 0$, we can write equality $\text{roots}(C_{row1}) = z_1$ and $\text{roots}(C_{row2}) = z_2$, but for $\alpha \neq 0$ $\text{angle}(\text{roots}(C_{row1})) = \text{angle}(z_1)$ and $\text{angle}(\text{roots}(C_{row2})) = \text{angle}(z_2)$. Based on these observations, regarding the assumptions of the model order, the second stage of Prony's method can be simplified. This simplification will enable to determine the initial stages of the components.

In the first stage, on the basis of a previously determined single zero position z_1 we determine the coefficients of a certain polynomial $\Psi(z)$ of first degree; let:

$$\Psi(z) = B_1 z + B_0. \tag{21}$$

Having the root z_1 of this polynomial, based on the relation describing the zero of the equation of the straight line:

$$z_1 = \frac{-B_0}{B_1} \tag{22}$$

we can write down:

$$\Psi(z) = B_1 z - B_1 z_1. \tag{23}$$

In the next step, we can create a vector with the coefficients of the searched FIR filter:

$$C_{row1} = [-B_1 z_1 \quad B_1]. \tag{24}$$

However, the parameter B_1 is still unknown. Exact knowledge of this parameter would enable estimation of the amplitude and phase of the desired components. It turns out, however, that to obtain the correct estimation of only the initial phase, it is enough to accept the value from the example shown, *i.e.*: 0–3.236 for $B1$ or another value with a zero real part and a negative imaginary part. For further transformations for simplicity, it was assumed that $B_1 = -i$. In this way, substituting $B_1 = -i$ in the formula (24) and next in the formula (20), and substituting $x_1 = a$ and $x_2 = b$ after transformations, it can be written down as a simple relation to the real and imaginary parts of the parameter h_1:

$$\text{Re}\{h_1\} = -\text{Im}\{z_1\} \cdot a, \tag{25}$$

$$\text{Im}\{h_1\} = \text{Re}\{z_1\} \cdot a - b. \tag{26}$$

By substituting the determined parameter in (19), a valid initial phase of a single component can be determined. The number of operations required for this purpose is small and is based on real data types. However, another solution should be used to determine the correct amplitude of Prony's model. The easiest way is to use the relation:

$$x_1 = amp \cdot \cos(2\pi f t_1 + \varphi). \tag{27}$$

If we make calculations for time $t_1 = 0$, with the substitution $x_1 = a$ we will obtain:

$$amp = \frac{a}{\cos(\varphi)}. \tag{28}$$

Ultimately, the calculation of the initial stage and the amplitude of Prony's model can be reduced to just a few lines of code in C or Matlab:

```
reh1=-imz1*a;
imh1=rez1*a-b;
absh1=sqrt(reh1*reh1+imh1*imh1);
FI=asin(imh1/absh1);
AMP=a/cos(FI).
```

The second stage of Prony's method requires a total of 4 multiplications, 2 additions, 2 divisions, and 1 root extraction operation, 1 arcsin and 1 cosine operation. All operations are performed on real data types.

3. Implementation of method in embedded device

The C-code method described in Section 2 was implemented in embedded devices with microprocessors of different architectures. These were the following microprocessors:

1) NUC140VE3CN in module Nuvoton Nu-LB-NUC140 [22];
2) LM4F120H5QR in module Stellaris LM4F120 LaunchPad Evaluation Kit [23];
3) TMS320C28027 in module C2000 Piccolo LaunchPad LAUNCHXL-F28027 [24];
4) MSP430G2553 in module MSP-EXP430G2 TI LaunchPad [25].

The execution times of the method were analysed for selected hardware platforms. The results are shown in Table 1. All calculations, except for the selected case, were made for 64-bit double numbers.

Table 1. Comparison of computation times for different processor systems.

Processor	clock [MHz]	computation time [μs]
LM4F120H5QR	80	215.1
TMS320C28027	60	275.4
NUC140VE3CN	50	337.1
MSP430G2553 (32bit)	16	1161.1
MSP430G2553 (64bit)	16	3733.0

The execution times of individual instruction groups were also analysed in relation to the execution time of the entire algorithm. The analysis was performed for the Nuvoton Nu-LB-NUC140 platform and the results are presented in Table 2.

Table 2. A summary of the numbers of operations and their execution times by the original Prony's method with $p = 2$, for Nuvoton Nu-LB-NUC140, with $f_{CLK} = 50$ MHz.

operation	number of operations I stage	number of operations II stage	operations in total	execution time 1 instruction [μs]	execution time instructions [μs]	instruction percentage
*	24	4	28	4.2	117.6	35%
+	12	2	14	2.3	32.2	10%
/	5	2	7	7.7	53.9	16%
√	2	1	3	11.7	35.1	10%
arcsin	1	1	2	25.0	50.0	15%
cos	0	1	1	25.0	25.0	7%
ln	1	0	1	23.3	23.3	7%
total:	**45**	**11**	**56**		**337.1**	**100%**

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 721–728.

4. Conclusions

A modification of Prony's method presented in the paper enables simple implementation of the original Prony's second-order method in embedded devices with a low computing power. The optimization of the algorithm enabled to use short algorithms in a wide range of measurement devices that perform measurement of a single sine component of an analysed signal and in security devices that have a short response time to specific events. Because the order of the model is limited in the algorithm to one real component for practical applications in which the analysed signal contains different distortions, the best results of the method can be achieved using an additional bandpass filter in the algorithm prior to the analysis by Prony's method [26]. The basic parameters of the bandpass filter, such as bandwidth, waving, and stopband damping should be chosen in accordance with a specific implementation. Examples of implementation of the method for specific measurement applications along with the selection of a bandpass filter will be the subject of further publications.

References

[1] Duda, K., Zieliński, T.P., Magalas, L.B., Majewski, M. (2011). DFT based Estimation of Damped Oscillation's Parameters in Low-frequency Mechanical Spectroscopy. *IEEE Trans. Instrum. Meas.*, 60(11), 3608–3618.

[2] Duda, K. (2011). DFT interpolation algorithm for Keiser-Bessel and Dolph-Chebyshev windows. *IEEE Trans. Instrum. Meas.*, 60(3), 784–790.

[3] Yu, C., Huang, Y., Jiang, J. (2010). A Full- and Half- Cycle DFT-based Technique for Fault Current Filtering. *2010 IEEE International Conference on Industrial Technology (ICIT), Vina del Mar*, *Chile*, 14–17.

[4] Wu, R.C., Chiang, C.T. (2010). Analysis of the Exponential Signal by the Interpolated DFT Algorithm. *IEEE Trans. Instrum. Meas.*, 59(12), 3306–3317.

[5] Borkowski, J., Mroczka, J. (2010). LIDFT method with classic data windows and zero padding in multifrequency signal analysis. *Measurement (London)*, 43(10), 1595–1602.

[6] Borkowski, J., Mroczka, J. (2002). Metrological analysis of the LIDFT method. *IEEE Transactions on Instrumentation and Measurement*, 51(1), 67–71.

[7] Wen, H., Teng, Z., Wang, Y., Zeng, B., Hu, X. (2011). Simple Interpolated FFT Algorithm Based on Minimize Sidelobe Windows for Power-Harmonic Analysis. *IEEE Trans. Power Electronics*, 26(9), 2570–2579.

[8] Belega, D., Petri, D. (2013). Accuracy Analysis of the Multicycle Synchrophasor Estimator Provided by the Interpolated DFT Algorithm. *IEEE Trans. Instrum. Meas.*, 62(5), 942–953.

[9] Borkowski, J.S., Kania, D.Ł., Mroczka, J. (2014). Influence of A/D quantization in an interpolated DFT based system of power control with a small delay. *Metrol. Meas. Syst.*, 21(3), 423–432.

[10] Szmajda, M., Górecki, K., Mroczka, J. (2010). Gabor Transform, SPWVD, Gabor-Wigner Transform and Wavelet Transform. *Tools For Power Quality Monitoring*, 17(3), 383–396.

[11] Delfino, F., Procopio, R., Rossi, M., Rachidi, F. (2012). Prony Series Representation for the Lightning Channel Base Current. *IEEE Trans. Electromagnetic Compatibility*, 54(2), 308–315.

[12] Peng, J.C. H., Nair, N.K.C. (2009). Adaptive sampling scheme for monitoring oscillations using Prony analysis. *Generation, Transmission & Distribution*, *IET*, 3(12), 1052–1060.

[13] Tawfik, M.M., Morcos, M.M. (2005). On the use of Prony method to locate faults in loop systems by utilizing modal parameters of fault current. *IEEE Trans. Power Del.*, 20(1), 532–534.

[14] Tawfik, M.M., Morcos, M.M. (2006). Fault Location on Loop Systems Using the Prony Algorithm. *Electric Power Components and Systems*, 34(4), 433–444.

[15] Zahlay, F.D., Rama Rao, K.S. (2012). Neuro-Prony and Taguchi's methodology based adaptive autoreclosure scheme for electric transmission systems. *IEEE Trans. Power Del.*, 27(2), 575–582.

[16] Zygarlicki, J., Mroczka, J. (2014). Prony's method with reduced sampling – numerical aspects. *Metrol. Meas. Syst.*, 21(3), 521–534.

[17] Zygarlicki, J., Zygarlicka, M., Mroczka, J., Latawiec, K. (2010). A reduced Prony's method in power quality analysis – parameters selection. *IEEE Transactions on Power Delivery*, 25(2), 979–986.

[18] Zygarlicki, J., Mroczka, J. (2012). Variable-frequency Prony method in the analysis of electrical power quality. *Metrol. Meas. Syst.*, 19(1), 39–48.

[19] Zygarlicki, J., Mroczka, J. (2012). Prony method used for testing harmonics and interharmonics of electric power signals. *Metrol. Meas. Syst.*, 19(4), 659–672.

[20] Zygarlicki, J., Zygarlicka, M., Mroczka, J. (2008). Prony's method in power quality analysis. *Proc. 9th Int. Scientific Conf. Electric Power Engineering (EPE), Brno, Czech Republic*, 115–119.

[21] Marple, S., Lawrence, J. (1987). *Digital Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

[22] *https://www.nuvoton.com/resource-files/DA00-NUC140ENF1.pdf* (May 2017).

[23] *http://www.ti.com/lit/ds/symlink/tm4c1233h6pm.pdf* (May 2017).

[24] *http://www.ti.com/lit/ds/symlink/tms320f28027.pdf* (May 2017).

[25] *http://www.ti.com/lit/ds/slas735j/slas735j.pdf* (May 2017).

[26] Kumaresan, R., Feng, Y. (1991). FIR prefiltering improves Prony's method. *IEEE Trans. Signal Processing*, 39(3), 736–741.

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 729–738.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

THEORETICAL SIMULATION OF A ROOM TEMPERATURE HgCdTe LONG-WAVE DETECTOR FOR FAST RESPONSE – OPERATING UNDER ZERO BIAS CONDITIONS

Piotr Martyniuk, Małgorzata Kopytko, Paweł Madejczyk, Aleksandra Henig, Kacper Grodecki, Waldemar Gawron, Jarosław Rutkowski

Military University of Technology, Institute of Applied Physics, Gen. S. Kaliskiego 2, 00-908 Warsaw, Poland
(✉ piotr.martyniuk@wat.edu.pl, +48 26 183 9215, malgorzata.kopytko@wat.edu.pl, pawel.madejczyk@wat.edu.pl, aleksandra.henig@wat.edu.pl, kacper.grodecki@wat.edu.pl, wgawron@vigo.com.pl, jaroslaw.rutkowski@wat.edu.pl)

Abstract

The paper reports on a long-wave infrared (cut-off wavelength ~ 9 μm) HgCdTe detector operating under unbiased condition and room temperature (300 K) for both short response time and high detectivity operation. The optimal structure in terms of the response time and detectivity versus device architecture was shown. The response time of the long-wave (active layer Cd composition, $x_{Cd} = 0.19$) HgCdTe detector for 300 K was calculated at a level of $\tau_s \sim 1$ ns for zero bias condition, while the detectivity – at a level of $D^* \sim 10^9$ cmHz$^{1/2}$/W assuming immersion. It was presented that parameters of the active layer and P^+ barrier layer play a critical role in order to reach $\tau_s \leq 1$ ns. An extra series resistance related to the *processing* (R_{S+} in a range 5−10 Ω) increased the response time more than two times ($\tau_s \sim 2.3$ ns).

Keywords: response time, unbiased condition, HgCdTe, LWIR, higher operating temperature.

1. Introduction

The room operating temperature condition of long-wave (8−12 μm, LWIR) detectors is the key research area in the infrared technology covering many applications [1]. Those new applications directly contribute to stringent (especially for zero bias and room temperature $T \sim 300$ K) operating conditions – a short response time (< 1 ns) and a high detectivity corresponding to the background limited photodetector – BLIP (> 10^9 cmHz$^{1/2}$/W) [2]. The requirement of the zero bias structures is mostly related to the fact that biased structures operating under non-equilibrium conditions exhibit an increase of $1/f$ noise, while the lack of cooling reduces the cost of the devices [3]. According to the experimental results presented in our previous papers, the LWIR HgCdTe $N^+\pi P^+n^+$ (π stands for a low doped p-type active layer) photodetectors operating under non-equilibrium conditions and room temperature reach a response time in a several nanosecond range and an optimized detectivity (p type absorber's doping < 10^{16} cm^{-3}) – $D^* \sim 10^{10}$ cmHz$^{1/2}$/W (assuming that the detector is immersed) [4−8].

For unbiased structures a fast loss in a signal due to a high rate of recombination in the absorber region and a very short passage time of carriers through the depletion area play the decisive role in ultra-fast response operation. To fabricate a detector exhibiting both short response time and high detectivity operating under zero bias conditions, multilayer hetero-structures must be implemented. In terms of a short response time, the p-type absorber is more useful due to high carrier ambipolar mobility. Additionally, in terms of reaching a high detectivity, in the fundamental approach, the p-type HgCdTe active regions exhibit the best

Article history: received on Apr. 28, 2017; accepted on Jun. 24, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0055.

compromise between the requirement of high quantum efficiency and a low thermal *generation-recombination* (GR) rate [9]. The mentioned complex multi-layer structures should consist of proper doping and composition gradient layers. Our main contribution to the field in comparison with the well-known three-layer N^+pP^+ structure invented and introduced by Elliot *et al.* for non-equilibrium conditions is intentional doping and composition gradient layers at heterojunctions (interfaces) [10−17].

The paper deals with the theoretical simulation of a photodetector for fast response conditions (< 1 ns) based on epitaxial HgCdTe multi-layer graded gap architecture. The detector structure was simulated with APSYS software by Crosslight Inc. [18]. A time response of the LWIR HgCdTe detector with the active layer composition $x_{Cd} = 0.19$ at $T = 300$ K was estimated at a level of $\tau_s \sim 1$ ns for a series resistance $R_S = 0.77$ Ω (an extra series resistance $R_{S+} = 0$ Ω). It was shown that the extra series resistance, related to the *processing* (R_{S+} in the range 5−10 Ω) increased the response time more than two times ($\tau_s \sim 2.3$ ns) for zero bias condition.

2. Architecture for unbiased condition

A graph of the simulated multilayer $N^+pP^+n^+$ structure for unbiased condition and operating in room temperature exhibiting both short response time and high detectivity is presented in Fig. 1, where detailed input parameter values of included gradient layers (interface layers) with proper doping and composition gradients are shown. Other parameters taken in modelling of LWIR detector are presented in Table 1. As mentioned above, the highly doped *p*-type active layer will be preferable to meet the requirement of a short time response. Both N^+ and P^+ wide bandgap barrier layers should be highly doped ($N_D, N_A > 10^{17}$ cm^{-3}) to minimize the diffusion length for carriers generated in regions close to the electric contacts and device resistance. Additionally, a highly doped n^+ ($N_D > 10^{17}$ cm^{-3}) layer provides a low resistance contact to metallization. In addition, the N^+ contact/barrier layer plays also the role of a radiation window and should be chosen in relation to absorber in terms of x_{Cd} to determine an appropriate cut-on wavelength of the photodetector.

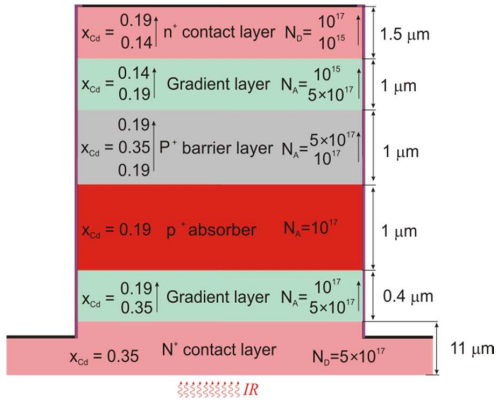

Fig. 1. The simulated HgCdTe multilayer $N^+pP^+n^+$ structure. Layers' N_D, N_A doping in cm^{-3}.

Optimization regarding a short response time and high detectivity is difficult due to the fact that both parameters rule out one another. Since the detector operates under zero bias, it was assumed that detectivity was limited by thermal Johnson-Nyquist noise. In addition, the structure could be optimized in terms of a short response time without a significant reduction in detectivity by assuming the immersion effect increasing detectivity by $\sim n^2$, where n stands for the GaAs substrate refractive index, according to the relation:

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 729–738.

$$D^* = \frac{n^2 R_i}{(4k_B T/R_o A)^{0.5}}, \tag{1}$$

where: R_i, k_B, R_o, A stand for current responsivity, the Boltzmann constant, a resistance at zero bias and the detector's electrical area, respectively. In terms of a response time the photocurrent dependence on time was used, where time for a $1/e$ drop from the photocurrent's maximum value was assessed.

The dependence of the response time and detectivity on the complex multilayer architecture, *i.e.* x_{Cd} composition, d thickness, N_A, N_D doping of a single layer and doping gradients, was simulated. In addition, the dependencies of the structure resistance and capacitance on the structural parameters were presented. The numerical simulation of the HgCdTe multilayer hetero-structure was performed by APSYS platform (Crosslight Inc.). The numerical model implemented in APSYS platform incorporates HgCdTe GR mechanisms. The absorption's coefficient (α), temperature, doping and composition dependence were assumed (for $T = 300$ K and active layer composition, $x_{Cd} = 0.19$, $\alpha = 137923$ m^{-1}). All equations describing GR models and parameters as intrinsic concentration, bandgap energy, carrier mobility, dielectric constants used in simulations were taken after the monograph by Capper and the APSYS manual [18–20]. The model given by Li *et al.* used in response time calculation and simulation was performed for $\lambda = 8$ μm [21].

Table 1. The parameters taken in modelling of LWIR detector.

Parameter	Symbol	Value
N_D, N_A – doping gauss tail	dx [μm]	0.02
trap density	N_T [cm^{-3}]	10^{14}
trap energy level	E_T	1/3 E_g (with respect to E_c)
capture coefficients SRH recombination centers	c_n, c_p [cm^3 s^{-1}]	1.5×10^{-7}, 3×10^{-9}
detector's electrical area	A [μm^2]	100×100
overlap integrals for Bloch functions	F_1F_2	0.3
incident power density	P [W m^{-2}]	500

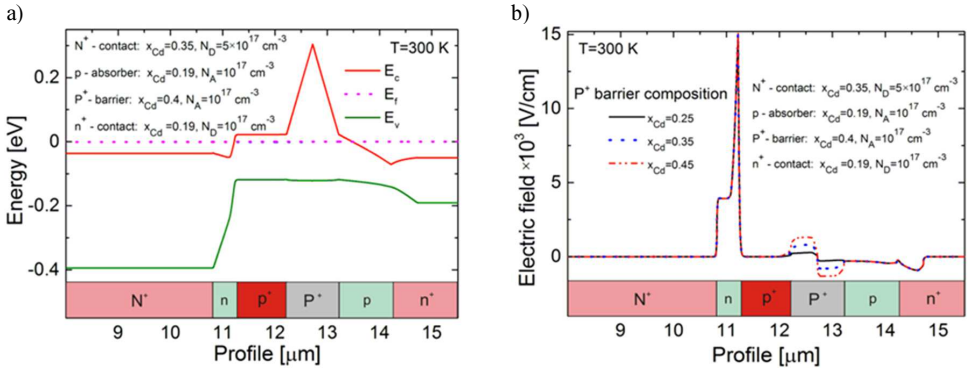

Fig. 2. The energy band profile (a) and the electric field (b) at HgCdTe multilayer N$^+$pP$^+$n$^+$ structure calculated for $T = 300$ K; N^+ contact $x_{Cd} = 0.35$, $N_D = 5\times10^{17}$ cm^{-3}; p-type absorber $x_{Cd} = 0.19$, $N_A = 10^{17}$ cm^{-3}; P$^+$ barrier $x_{Cd} = 0.25$–0.45, $N_A = 10^{17}$ cm^{-3}; n$^+$ contact $x_{Cd} = 0.19$, $N_D = 10^{17}$ cm^{-3}.

The energy band profile calculated for the p type active layer $N_A = 10^{17}$ cm^{-3} ($N_A > 3n_i \sim 3.6 \times 10^{16}$ cm^{-3}) and $T = 300$ K is presented in Fig. 2a. Composition and doping gradients at the interface between the N$^+$ contact and p-type doped active layers were chosen in order not

to generate discontinuities in the energy band profile among main constituent heterojunctions of the $N^+pP^+n^+$ structure. The electric field drops mostly on that interface ($E \sim 15 \times 10^3$ V/cm), which is shown in Fig. 2b. In addition, the electric field dependence on P^+ barrier composition ($x_{Cd} = 0.25–0.45$) is also shown pointing out that the P^+ layer x_{Cd} marginally contributes to the electric field drop along the structure. It is clearly visible that for zero bias condition the electric field does not drop on the active layer, meaning that the response time is fundamentally limited by diffusion of the photo-generated carriers.

Both absorber's doping N_A and absorber's thickness d have a significant impact on the response time. The absorber's doping influence on the response time is presented in Fig. 3a. The absorber thickness $d = 1$ µm and no extra series resistance ($R_{S+} = 0$ Ω) were assumed in simulations. The response time decreases from ~ 1.8 ns to ~ 1 ns for the absorber's doping $N_A = 5 \times 10^{15}–10^{17}$ cm^{-3}. To compare those results with the diffusion time we assume that an electron to hole mobility ratio ~ 100 (for $x_{Cd} = 0.19$, $N_A = 5 \times 10^{15}$ cm^{-3} an ambipolar diffusion coefficient and ambipolar diffusion length are $D_a = 5.6 \times 10^{-4}$ m^2/s and $L_a = 2.16$ µm, respectively, while for $N_A = 10^{17}$ cm^{-3} – they are $D_a = 2.73 \times 10^{-3}$ m^2/s, $L_a = 4.08$ µm, for $T = 300$ K). The diffusion time is given by – $\tau_{diff} = \frac{d^2}{2.4D_a}$ for $d << L_a$ and changes from 0.74 ns to 0.15 ns for $N_A = 5 \times 10^{15}–10^{17}$ cm^{-3}. That confirms that highly doped p-type material significantly reduces the diffusion time. Another factor contributing to the response time is the RC constant τ_{RC}. The structure capacitance C and series resistance R_S, presented in Fig. 3b, were assessed at levels of ~ 0.18 nF and 0.77 Ω, resulting in extra $\tau_{RC} \sim 0.14$ ns for $N_A = 10^{17}$ cm^{-3} and $\tau_{RC} \sim 0.19$ ns for $N_A = 5 \times 10^{15}$ cm^{-3}, respectively. The diffusion time τ_{diff} and time constant τ_{RC} are shown in Fig. 3a and in the whole absorber's doping range they have smaller values than the response time.

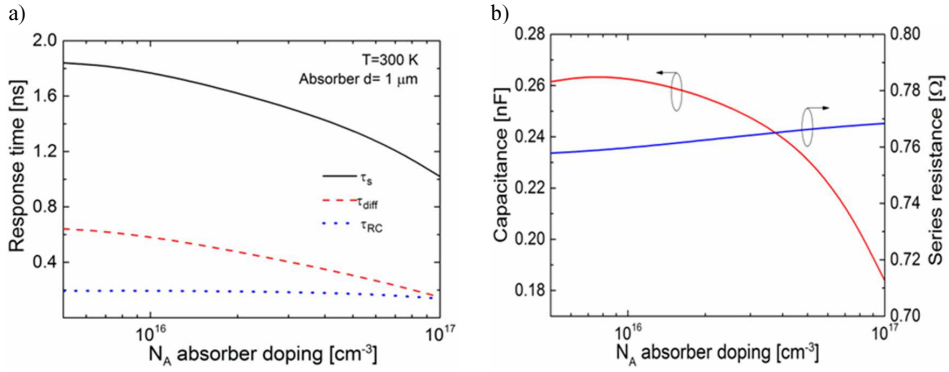

Fig. 3. Response time, diffusion time and RC constant calculated for $T = 300$ K versus absorber's doping $N_A = 5 \times 10^{15}–10^{17}$ cm^{-3} (a). Structure capacitance and resistance versus absorber's doping (b).

The diffusion time depends on the distance range between the generation and depletion areas, meaning that the response time will be suppressed with a reduction of active layer thickness. Assuming the absorber's thickness $d = 1–5$µm, active layer doping $N_A = 10^{17}$ cm^{-3} and no extra resistance contribution ($R_{S+} = 0$ Ω), the response time stays within a range $\tau_s \sim 1–3.08$ ns but the diffusion time reaches $\tau_{diff} = 0.15–2.35$ ns, as it is shown in Fig. 4a. The RC constant calculated for the data presented in Fig. 4 (b) reaches $\tau_{RC} \sim 0.14–0.24$ ns for $d = 1–5$ µm, respectively.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 729–738.

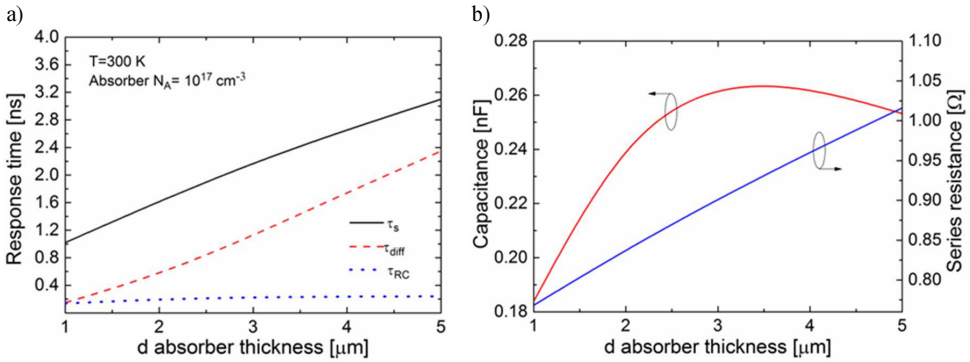

Fig. 4. Response time, diffusion time and RC constant calculated for $T = 300$ K versus absorber's thickness $d = 1–5$ μm for absorber's doping $N_A = 10^{17}$ cm^{-3} (a). Capacitance and resistance versus absorber d (b).

The dependencies of responsivity and detectivity on absorber's doping and absorber's thickness are presented in Fig. 5. The maximum detectivity was found at the level of ~ 2.5×10^9 cmHz$^{1/2}$/W for the absorber doping $N_A = 10^{17}$ cm^{-3} (that nearly corresponds to $N_A \sim 3n_i$ being an optimal value where Auger thermal generation reaches its minimal value) for $d = 1$ μm and for $T = 300$ K, what is shown in Fig. 5a. The maximum detectivity increases with absorber's thickness and reaches its maximum value for absorber's thickness $d \sim 3.5$ μm, what is presented in Fig. 5b. The response time at a level of 1 ns can be obtained for the optimum absorber's thickness $d = 1$ μm and absorber's doping $N_A = 10^{17}$ cm^{-3} with detectivity greater than ~ 2.5×10^9 cmHz$^{1/2}$/W.

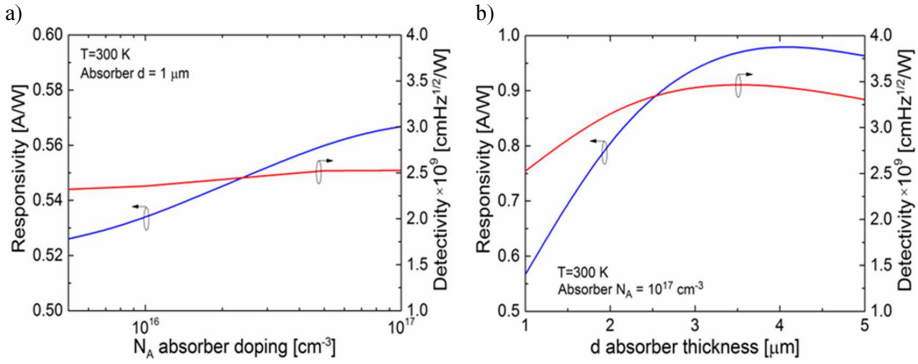

Fig. 5. Responsivity and detectivity versus absorber's doping for absorber's thickness $d = 1$ μm (a). Responsivity and detectivity versus absorber's thickness, absorber's doping $N_A = 10^{17}$ cm^{-3} (b).

In the whole simulation of the absorber's contribution, the response time was higher than the diffusion time and RC constant (see Figs. 3a and 4a). In order to clarify these differences an impact of other constituent layers of the heterojunction on the photodetector's response time was analysed. The dependence of the response time and detectivity on the N^+ contact layer thickness within a range $d = 2–20$ μm was calculated. The N^+ contact thickness marginally contributes to the response time. For $d > 9$ μm the response time saturates reaching 0.98 ns, what is shown in Fig. 6a. The shorter N^+ contact layer the slightly higher response is reached ($\tau_s \sim 1.34$ ns was assessed for $d = 2$ μm). Within a range $d = 1.5–9$ μm of the N^+ layer thickness, the response time could be increased by 350 ps, what is shown in Fig. 6a. D^* also stays constant

within the analysed N^+ contact layer thickness reaching ~ 2.5×10^9 cmHz$^{1/2}$/W, what is presented in Fig. 6b.

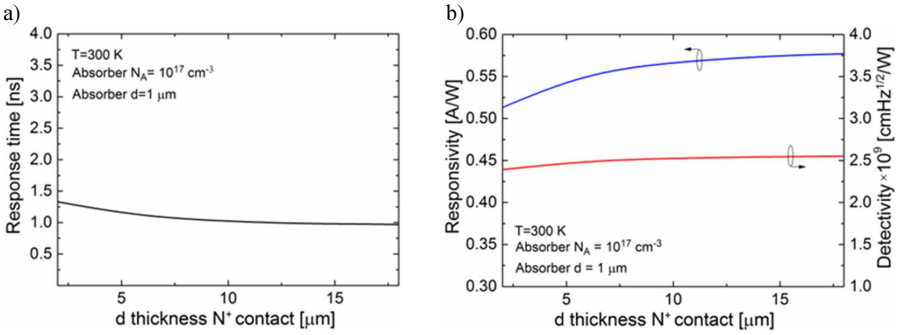

Fig. 6. Response time calculated for $T = 300$ K versus N^+ thickness contact layer $d = 2-20$ μm, absorber's doping $N_A = 10^{17}$ cm^{-3} (a). Responsivity and detectivity versus N^+ thickness contact layer, absorber's thickness $d = 1$ μm, absorber's doping $N_A = 10^{17}$ cm^{-3} (b).

The N^+ contact layer must be transparent for IR and must be chosen properly in relation to the active layer regarding its composition and thickness (resistance – mesa structure). Doping must be high $N_D > 10^{17}$ cm^{-3} to make the metallization to that layer. The higher composition the slightly higher response time is reached, however x_{Cd} marginally contributes to τ_s, meaning that lowering x_{Cd} to a level slightly higher than the active layer enables to decrease τ_s by ~ 220 ps for the N^+ contact layer doping $N_D = 5 \times 10^{17}$ cm^{-3}, what is presented in Fig. 7a. The N^+ layer x_{Cd} has also a marginal influence on detectivity reaching ~ 2.5×10^9 cmHz$^{1/2}$/W, what is shown in Fig. 7b.

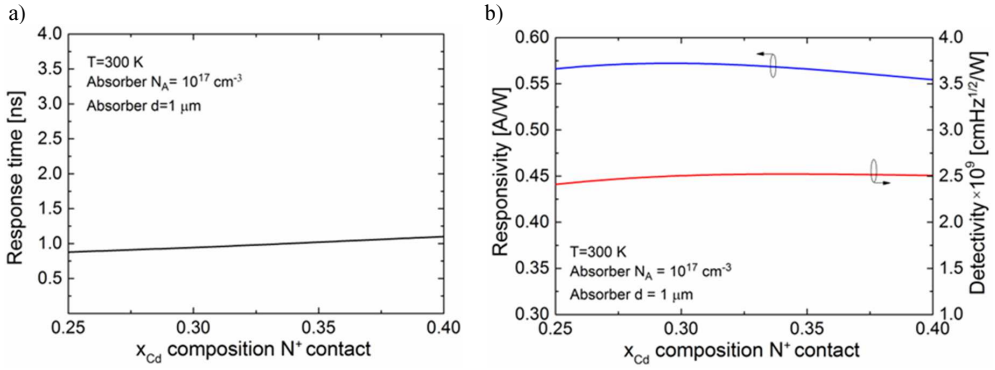

Fig. 7. Response time calculated for $T = 300$ K versus N^+ contact layer composition $x_{Cd} = 0.25-0.4$, absorber's doping $N_A = 10^{17}$ cm^{-3} (a). Responsivity and detectivity versus N^+ contact layer x_{Cd} composition, absorber's doping $N_A = 10^{17}$ cm^{-3} (b).

The contribution of the P^+ barrier composition within a range $x_{Cd} = 0.2-0.45$ to τ_s and D^* was simulated. The P^+ barrier composition should be at a level of $x_{Cd} < 0.35$ to reach an ultra-short response time $\tau_s < 1$ ns for an extra series resistance $R_{S+} = 0$ Ω, what is presented in Fig. 8a. P^+ x_{Cd} has a significant influence on detectivity, what is shown in Fig. 8b. For P^+ $x_{Cd} < 0.4$ the detectivity D^* decreases rapidly for analysed conditions, its maximum value being

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 729–738.

D^*~ 3.3 × 10^9 cmHz$^{1/2}$/W for x_{Cd} = 0.45, whereas for the x_{Cd} comparable to the active layer D^* reaches ~ 1.25 × 10^9 cmHz$^{1/2}$/W assuming R_{S+} = 0 Ω.

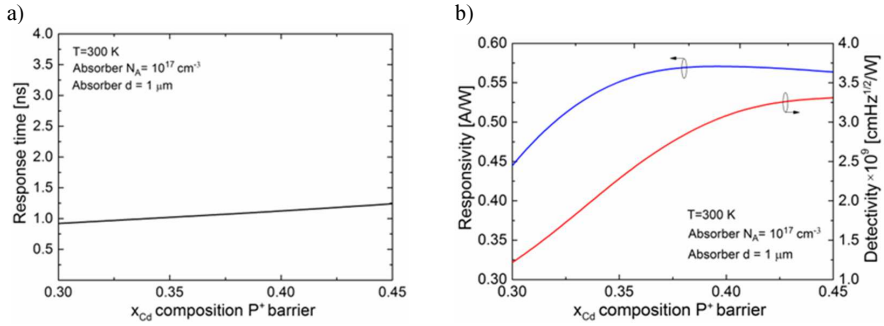

Fig. 8. Response time calculated for T = 300 K versus P$^+$ barrier layer composition x_{Cd} =0.3−0.45, absorber's doping N_A = 10^{17} cm^{-3} (a). Responsivity and detectivity versus P$^+$ barrier layer x_{Cd} composition, absorber's doping N_A = 10^{17} cm^{-3} (b).

The influence of the P$^+$ barrier layer doping within a range N_A ~ 5 × 10^{15}–5 × 10^{17} cm^{-3} on τ_s and D^* is presented in Figs. 9a and 9b. The shortest response time τ_s ~ 0.7 ns was reached for N_A ~ 5 × 10^{17} cm^{-3} doping being higher than doping of the absorber layer (N_A = 10^{17} cm^{-3}). The P$^+$ barrier doping highly influences both responsivity and detectivity characteristics. Once the P$^+$ barrier doping increases within a range N_D = 5 × 10^{15} – 5 × 10^{17} cm^{-3}, the detectivity changes from ~ 7.6 × 10^8 to ~3.3 × 10^9 cmHz$^{1/2}$/W. The optimal value of P$^+$ barrier layer doping should be greater than N_A = 10^{17} cm^{-3}.

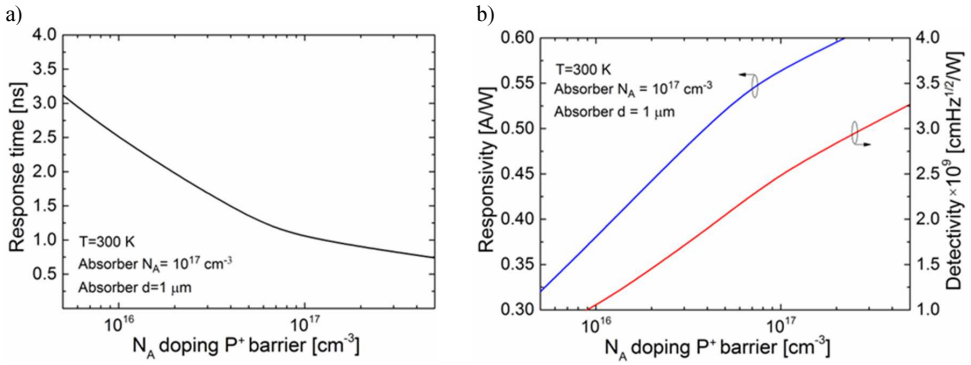

Fig. 9. Response time calculated for T = 300 K versus P$^+$ barrier layer doping N_A = 5×10^{15}–5×10^{17} cm^{-3}, absorber's doping N_A = 10^{17} cm^{-3} (a). Responsivity and detectivity versus P$^+$ barrier layer doping, absorber's doping N_A = 10^{17} cm^{-3} (b).

The influence of the P$^+$ barrier layer thickness within a range d = 0.5−3 μm on τ_s and D^* was simulated. Increasing the P$^+$ barrier width causes a very small increase in the response time and has also a small effect on both response time and detectivity, what is shown in Figs. 10a and 10b. The response time changes from to ~ 1 to ~ 1.1 ns for d = 0.5−3 μm. The response time $\tau_s \leq$ 1 ns could be reached for the P$^+$ barrier thickness d < 1 μm, for the absorber's doping N_A = 10^{17} cm^{-3} assuming no extra R_{S+}. D^* stays constant within the analysed P$^+$ barrier layer range reaching ~ 2.5 × 10^9 cmHz$^{1/2}$/W.

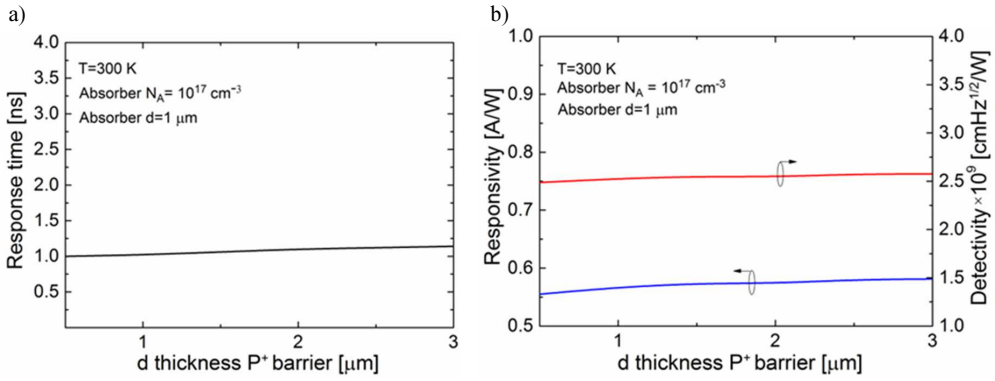

Fig. 10. Response time calculated for T = 300 K versus P$^+$ barrier thickness d = 0.5−3 μm, absorber's doping N_A = 10^{17} cm^{-3} (a). Responsivity and detectivity versus P$^+$ barrier thickness d, absorber's doping N_A = 10^{17} cm^{-3} (b).

The influence of the thickness of *n*-type interface layer d = 0.4−2 μm between N$^+$ contact and absorber on both τ_s and D^* is presented in Figs. 11a and 11b. The *n*-type interface thickness d has almost any contribution to the response time. For 0.4 < d < 2 μm the response time increases only by ~ 130 ps, whereas D^* reaches ~ 2.5×10^9 cmHz$^{1/2}$/W in the analysed *n*-type interface thickness range.

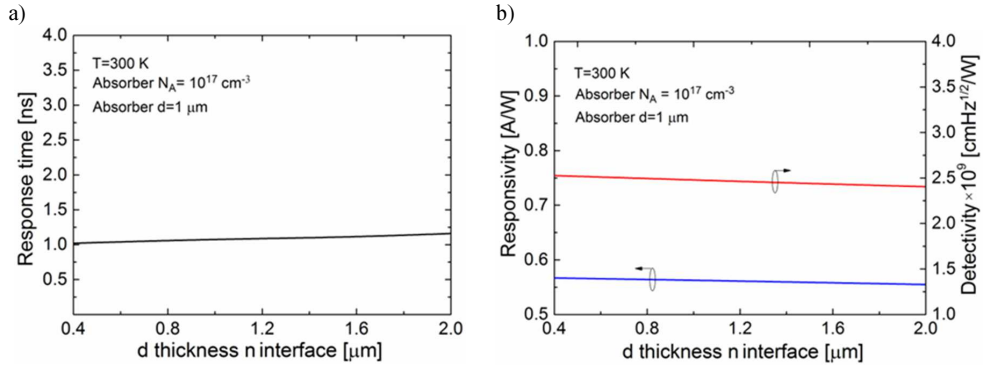

Fig. 11. Response time calculated for T = 300 K versus *n* type interface thickness d = 0.4−2 μm, absorber's doping N_A = 10^{17} cm^{-3} (a). Responsivity and detectivity versus *n* type interface thickness d, absorber's doping N_A = 10^{17} cm^{-3} (b).

The extra series resistance related to the processing, R_{S+}, plays a decisive role in increasing the response time. The absorber thickness d = 1 μm and extra series resistance from within a range R_{S+} = 0−10 Ω were assumed in simulations. The influence of the absorber's doping on the response time is presented in Fig. 12. The extra R_{S+}= 10 Ω increases the response time more than two times, i.e. for the absorber's N_A = 10^{17} cm^{-3} the response time changes within a range τ_s ~ 2.3−1 ns.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 729–738.

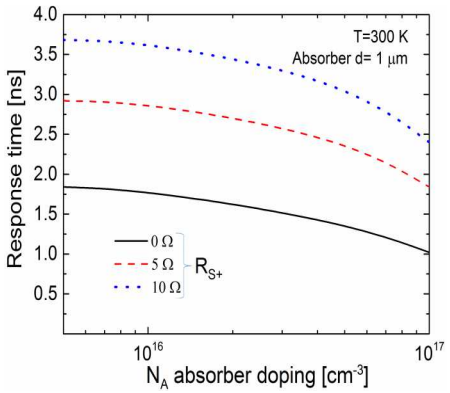

Fig. 12. Response time calculated for $T = 300$ K versus absorber's doping $N_A = 5\times10^{15}$–10^{17} cm^{-3}, extra $R_{S+} = 0$–10 Ω. Absorber's thickness $d = 1$ μm.

3. Conclusion

A structure operating in room temperature enabling to reach the response time $\tau_s < 1$ ns and detectivity $D^* > 2 \times 10^9$ cmHz$^{1/2}$/W was presented. The greatest impact on the response time is that of physical parameters of the active layer and P$^+$ barrier layer. Other layers have a marginal influence on both response time and detectivity. An optimal LWIR absorber $x_{Cd} = 0.19$ should meet following requirements: $N_A > 10^{17}$ and $d \leq 1$ μm. N$^+$ contact layer thickness $d \geq 10$ μm and $x_{Cd} < 0.35$. P$^+$ barrier layer $x_{Cd} \leq 0.35$, doping $N_A \geq 10^{17}$cm^{-3} and thickness d ≤ 2 μm. Assuming an extra series resistance $R_{S+} = 10$ Ω the response time increases to 2.3 ns.

Acknowledgments

We acknowledge the support of The National Centre for Research and Development – the grant no. TANGO1/2665576/NCBR/2015.

References

[1] Rogalski, A. (2011). *Infrared Detectors*. 2nd ed. CRC Press, Boca Raton.

[2] Wojtas, J., Bielecki, Z., Stacewicz, T., Mikołajczyk, J., Nowakowski, M. (2012). Ultrasensitive laser spectroscopy for breath analysis. *Opto-Electron. Rev.*, 20, 26–39.

[3] Elliot, C.T., Gordon, N.T., Hall, R.S., Phillips, T.J., Jones, C.L., Best, A. (1997). 1/*f* noise studies in uncooled narrow gap $Hg_{1-x}Cd_xTe$ non-equilibrium diodes. *J. Electron. Mater.*, 25, 643–648.

[4] Kopytko, M., Jóźwikowski, K., Madejczyk, P., Pusz, W., Rogalski, A. (2013). Analysis of the response time in high-temperature LWIR HgCdTe photodiodes operating in non-equilibrium mode. *Infrared Phys. Technol.*, 61, 162–166.

[5] Kopytko, M., Jóźwikowski, K., Rogalski, A., Jóźwikowska, A. (2010). High frequency response of near-room temperature LWIR HgCdTe heterostructure photodiodes. *Opto-Electron. Rev.*, 18, 277–283.

[6] Pawluczyk, J., Piotrowski, J., Pusz, W., Koźniewski, A., Orman, Z., Gawron, W., Piotrowski, A. (2015). Complex behavior of time response of HgCdTe HOT photodetectors. *J. Electron. Mater.*, 44, 3163–3173.

[7] Madejczyk, P., Gawron, W., Martyniuk, P., Kębłowski, A., Piotrowski, A., Pusz, W., Kowalewski, A., Piotrowski, J., Rogalski, A. (2013). MOCVD grown HgCdTe device structure for ambient temperature LWIR detectors. *Semicond. Sci. Technol.*, 28, 105017, 1–7.

[8] Madejczyk, P., Gawron, W., Martyniuk, P., Kębłowski, A., Pusz, W., Pawluczyk, J., Kopytko, M., Rutkowski, J., Rogalski, A., Piotrowski, J. (2017). Engineering steps for optimizing high temperature LWIR HgCdTe photodiodes. *Infrared Phys. Technol*., 81, 276–281.

[9] Rogalski, A. (2005). HgCdTe infrared detector material: history, status and outlook. *Rep. Prog. Phys.*, 68, 2267–2336.

[10] Ashley, T., Elliott, C.T. (1985). Non-equilibrium mode of operation for infrared detection. *Electron. Lett.*, 21, 451–452.

[11] Elliot, C.T., Gordon, N.T., Hall, R.S., Philips, T.J., White, A.M., Jones, C.L., Maxey, C.D., Metcalfe, N.E. (1996). Recent results on MOVPE grown heterostructure devices. *J. Electron. Mater.*, 25, 1139–1145.

[12] Emelie, P.Y., Philips, J.D., Velicu, S., Grein, C.H. (2007). Modeling and design consideration of HgCdTe infrared photodiodes under non equilibrium operation. *J. Electron. Mater.*, 36, 846–851.

[13] Emelie, P.Y., Velicu, S., Grein, C.H., Philips, J.D., Wijewarnasuriya, P.S., Dhar, N.K. (2008). Modeling of LWIR HgCdTe Auger-suppressed infrared photodiodes under non equilibrium operation. *J. Electron. Mater*., 37, 1362–1368.

[14] Piotrowski, A., Piotrowski, J., Gawron, W., Pawluczyk, J., Pędźinska, M. (2009). Extension of spectral range of Peltier cooled photodetectors to 16 μm. *Proc. SPIE*, 7298, 729824.

[15] Stanaszek, D., Piotrowski, J., Piotrowski, A., Gawron, W., Orman, Z., Paliwoda, R., Brudnowski, M., Pawluczyk, J., Pędzińska, M. (2009). Mid and long infrared detection modules for picosecond range measurements. *Proc. SPIE*, 7482, 74820M-74820M-11.

[16] Piotrowski, J., Pawluczyk, J., Piotrowski, A., Gawron, W., Romanis, M., Kłos, K. (2010). Uncooled MWIR and LWIR photodetectors in Poland. *Opto-Electron. Rev*., 18, 318–327.

[17] Velicu, S., Grein, C.H., Emelie, P.Y., Itsuno, A., Philips, J.D., Wijewarnasuriya, P. (2010). MWIR and LWIR HgCdTe infrared detectors operated with reduced cooling requirements. *J. Electron. Mater.*, 39, 873–881.

[18] APSYS Macro/User's Manual ver. 2011. (2011). Crosslight Software, Inc.

[19] Capper, P.P. *Properties of narrow gap cadmium-based compounds.* London, U.K.: Inst. Elect. Eng.

[20] Wenus, J., Rutkowski, J., Rogalski, A. (2001). Two-dimensional analysis of double-layer heterojunction HgCdTe Photodiodes. *IEEE Trans. Electron Devices*, 48, 7, 1326−1332.

[21] Li, Q., Dutton, R.W. (1991). Numerical small-signal AC modeling of deep-level-trap related frequency-dependent output conductance and capacitance for GaAs MESFET's on semi-insulating substrates. *IEEE Trans. Electron Devices*, 38, 1285–1288.

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

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS
Index 330930, ISSN 0860-8229
www.metrology.pg.gda.pl

M&MS

DEVELOPMENT OF A METHOD FOR TOOL WEAR ANALYSIS USING 3D SCANNING

Marek Hawryluk, Jacek Ziemba, Łukasz Dworzak

Wrocław University of Science and Technology, Faculty of Mechanical Engineering, I. Łukasiewicza 7-9, 50-371 Wrocław, Poland
(✉ marek.hawryluk@pwr.edu.pl, +48 71 320 2164, jacek.ziemba@pwr.edu.pl, lukasz.dworzak@pwr.edu.pl)

Abstract

The paper deals with evaluation of a 3D scanning method elaborated by the authors, by applying it to the analysis of the wear of forging tools. The 3D scanning method in the first place consists in the application of scanning to the analysis of changes in geometry of a forging tool by way of comparing the images of a worn tool with a CAD model or an image of a new tool. The method was evaluated in the context of the important measurement problems resulting from the extreme conditions present during the industrial hot forging processes. The method was used to evaluate wear of tools with an increasing wear degree, which made it possible to determine the wear characteristics in a function of the number of produced forgings. The following stage was the use it for a direct control of the quality and geometry changes of forging tools (without their disassembly) by way of a direct measurement of the geometry of periodically collected forgings (indirect method based on forgings). The final part of the study points to the advantages and disadvantages of the elaborated method as well as the potential directions of its further development.

Keywords: 3D scanning, measurements and volumetric analysis of tools wear, improve a durability of forging tools.

1. Introduction

The basis for the development of industry is the continuous improvement of products and their quality with simultaneous reduction of the production costs, which is indirectly associated with reducing the measurement time during very complicated measurement procedures. In the industrial coordinate metrology, new trends are observed, connected with the use of non-contact measurement techniques. It is possible owing to the continuous improvement of measurement techniques and introduction of new devices and measurement methods, such as fast scanning methods combined with CAD/CAM/FEM [9, 14, 18, 22, 27, 28]. It creates the necessity of applying numerical 3D models to the determination of nominal values during measurements. It is connected with the modern dimensional ISO GPS approach as well as with the integration of software for measurement appliances with numerical CAD models [15, 26, 28]. Additionally, due to the simplicity of their application and improvement of their measurement accuracy, mobile devices are objects of increasing interest of the industry. In reply to the market demand, one can observe continuous improving the measurement accuracy of industrial optical scanners as well as linear laser scanners, which, together with their mobility, significantly increases their competitiveness compared with the classic CMMs (*Coordinate Measuring Machines*) [2, 4, 12, 18, 32–35]. This is connected with the more and more frequent use of blue light sources instead of red light ones, which largely affects the measurement accuracy and – in the case of optical scanners - makes it possible to partly eliminate the necessity of detail matting. The more and more commonly used mobile devices include different types of optical scanners and measuring arms equipped with linear laser scanners with dedicated specialized software which, owing to their mobility and versatility, are an alternative for the coordinate measurement machines

Article history: received on Mar. 13, 2017; accepted on Jun. 24, 2017; available online on Dec. 01, 2017; DOI: 10.1515/mms-2017-0054.

in applications requiring lower accuracy [23]. For example, the accuracy of a mobile measuring arm was discussed in [29], whose authors performed tests consisting in evaluation of a representation of the nominal shape by means of an arm and a coordinate machine. The 3D scanning technique, also with the use of scanners, is mainly applied to the control of final quality of a product [30]. These measurements are usually based on evaluation of the form deviations of a selected contour and surface [16]. In many works much space is devoted to the use of CMMs and scanning methods to volumetric wear assessment of retrieved prostheses or bones [2, 3, 17, 19, 20, 25]. The available literature more and more often provides information on the application of this type of methods to the measurement, control and evaluation of the state of shaping tools. An example of such application of the 3D scanning method is the use of an optical scanner for determination of the form deviations of a selected surface and next, based on the obtained data, determination of the geometrical specification for the process of rebuilding [1, 24]. Another application of the 3D scanning methods with the use of scanners is the analysis of the form deviations of a selected surface to evaluate the wear of forging tools which are nitride covered or coated with hybrid layers. These analyses consist in comparison of the image obtained from scanning of a new forging tool before operation – with a reference CAD model and – next – with the image of the same tool after the forging, by way of determining the form deviations of the analysed surface [6, 7].

So far, it has been possible to find in the literature data concerning the use of 3D scanning methods only for analysis of the geometrical changes of a product (forgings) and possibly for the control of new tools, rather than for evaluation of states of the tools producing a given product or other applications of this type [13, 21]. This interest induces an analysis of scanning techniques regarding the possibility of their use and development in the forging industry, *e.g.* for analysis of the geometrical changes of tools during the forging process as well as for continuous evaluation of a forging tool based on periodically collected and scanned forgings, and more and more advanced analyses and applications.

The aim of the study is the elaboration and development of a non-contact measurement method – 3D scanning – for the analysis and evaluation of the wear of forging tools with the use of a measuring arm integrated with a linear laser scanner.

2. Measuring method and test bench description

In order to apply a non-contact method of scanning tools and forgings during the forging process, a ROMER Absolute ARM 7520si measuring arm was selected, together with an RS3 scanner (Fig. 1) and Polyworks software enabling to perform scanning in the Real-Time Quality Meshing technology.

Fig. 1. A Romer Absolute 7520si measuring arm.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

The selected arm makes it possible to perform contact measurements with the use of a contact measuring probe as well as non-contact measurements with the use of a linear laser scanner RS3 integrated with the arm.

The arm with the integrated scanner RS3 and the Real-Time Quality Meshing technology makes it possible to collect up to 460 000 points/s for 4600 points on the line with a linear frequency of 100 Hz, with a scanning system accuracy SI 0.058 mm in relation to B89.4.2. In order to perform the measurements, for the purposes of the elaborated measurement technology, laboratory test benches were constructed in a measuring laboratory, presented in Fig. 2.

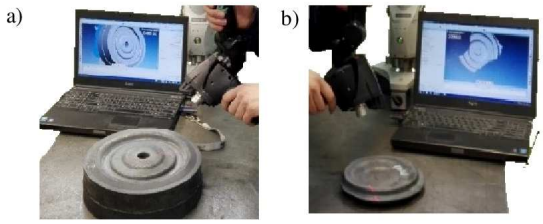

Fig. 2. The test bench for measurements with the use of a ROMER Absolute Arm 7520si measuring arm and an integrated laser scanner RS3, of: a die (a); a forging (b).

The test in relation to B89.4.2. consists in measuring a mat grey sphere by means of 5 different arm deviations. In each arm deviation, the sphere is scanned from 5 different directions, so that most of the sphere surface can be scanned. The result is the maximal 3D distance between the centre and the centres of 5 spheres. An error value obtained in this way is difficult to interpret in our measurement task.

Therefore, it was decided to perform additional accuracy tests that indicated the need to implement a software package – REAL-TIME QUALITY MESHING. This enables to eliminate defects of the linear scanner that are a result of the speed of movement and the position of the measuring head on the measurement accuracy. In the case when the scanning procedure is executed too fast or incorrectly, the model in the program shows unfilled "holes" or special coloured markers controlling the quality parameters of the scan. Such an approach to scanning with the use of a linear scanner, together with the application of the REAL-TIME QUALITY MESHING function, makes it possible to obtain a scan image with similar selected quality parameters.

The measurements with the use of non-contact measurement techniques, also with the application of the measuring arm with an integrated laser scanner, in the case of forging applications, are usually used for two types of objects: forgings and ready forged products, as well as forging tools and instrumentation. In the case of mobile measuring devices (measuring arms and scanners), much more popular are the measurements of forging instrumentation, owing to their mobility and capability of measuring large sized, heavy dies, often directly on the production line.

3. Description of technology

Based on numerous studies and the authors' experience, the development and evaluation of a 3D scanning method (with the use of non-contact techniques) for forging applications were presented, referring mainly to the measurements of the wear progress of forging tools. Fig. 3 shows a block diagram of the proposed method.

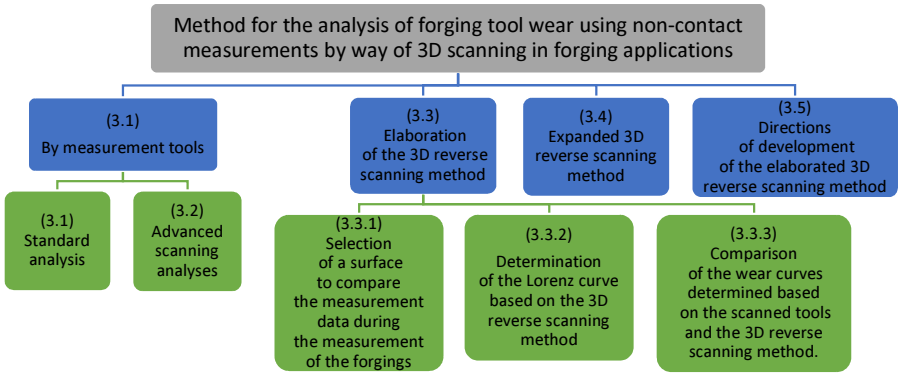

Fig. 3. A block diagram of evaluation of the method of forging tool wear analysis.

The proposed technology of evaluating tool wear based on the diagram presented in Fig. 3 is accomplished in the following stages:

3.1 Standard analysis. Performing a simple analysis of the material loss at the end of a tool's work in order to determine the general (typical) wear of its working surfaces.
- **3.1.1** Selection of reference surfaces for data equalization by the best-fit method.
- **3.1.2** Comparison of tools' images after operation (forging) with either the image of a new (unworn) tool or a CAD model (in the case of difficulty in scanning a new tool, *e.g.* due to deep working patterns).

3.2 Advanced scanning analyses, *e.g.* of the tool wear progress during forging. Determination of the Lorenz curve.

3.3 Elaboration of a 3D reverse scanning method. Analysis of tool wear (material loss) based on periodically collected forgings and measurement of the changes in their surfaces (material growth)
- **3.3.1** Selection of a surface of minor wear from the forging pattern to compare the measurement data during the measurement of the forgings.
- **3.3.2** Determination of the Lorenz curve based on the 3D reverse scanning method.
- **3.3.3** Comparison of the wear curves determined based on the scanned tools and the 3D reverse scanning method.

3.4 Expanded 3D reverse scanning method. Division of the tool into selected areas, according to the occurrence of different degradation mechanisms in these areas.
- **3.4.1** Limitations of the method (temperature, scale, different closings)

3.5 Directions of future development of the elaborated 3D reverse scanning method.

3.1. Standard analysis

The most commonly applied method of analysis by way of scanning is the standard analysis of the material at the end of the tool's work in order to determine the image of degradation of the working tool surface, typical for a given type of insert. The die insert wear is analysed using the data obtained from the measurements made by way of scanning the die before and after the operation, on the test bench shown in Fig. 2a. The data are compared by the GOM software with the function of best-fit data equalization.

The result of the analysis performed with the GOM software is a coloured map of deviations distributed on the scan surface, showing the value of the form deviation from the nominal dimension, that is the scan of the die before the forging process. The measurement results are presented in Fig. 4. Sometimes, in order to perform a simple and fast process analysis,

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

especially for axis-symmetrical tools, visualization of wear is applied on a selected tool section, with an additionally introduced scale of magnitude of the form deviation, *e.g.* 10 times (Fig. 4b).

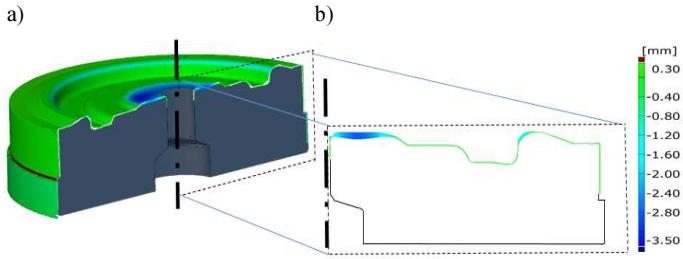

Fig. 4. An example of analysis results: a map of deviations for a die insert (a); scanning results for a selected longitudinal die section (b).

Figure 4 shows a symmetrical ring-shaped wear of the insert in its central part. In this area, radial grooves are visible on the deep ring; the maximal value of wear of this surface equals 3.17 mm. In turn, in the insert bridge area, asymmetrical ring-shaped wear can be seen, with the maximal value of 2.23 mm. No geometrical loss in the other areas is visible in the scan image.

3.1.1. Selection of reference surface for data comparison by best-fit method

A very important aspect is the selection of reference surfaces for data equalization by the best-fit method, in order to minimize errors during the analysis (Fig. 5). It is a crucial aspect in light of the applied mathematical algorithm and the analysis of details concerning the form deviation of a determined surface, often even at a level of a few millimetres, similarly to the case of tools used for preliminary forging operations. The authors' experience, confirmed by the studies, points to the surfaces not participating in the forging process as the ones most often selected for a reference. It should be noted that such an ideal solution is possible only in selected applications.

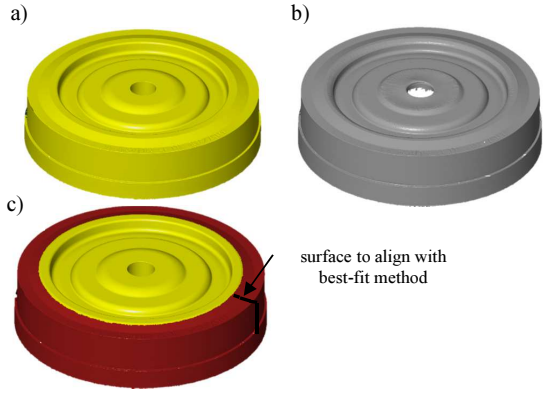

Fig. 5. An insert scan before forging (a) after forging (b) elements selected for the best-fit align with GOM software from the insert scan before forging (c).

For the tool shown in Fig. 4, such a surface is the area marked in red (Fig. 5c), for which it was established that, even with a very big number of forgings, its wear is scant due to their not taking part in the forging process. In the case when it is difficult to find or point to such surfaces which would be ideal for the equalization process, there are selected those tool surfaces for

which the wear in the forging process is scant. Of course, at times, the situation is much more complicated, as in the case of a punch-tool used in the second, upper operation of forging a lid (Fig. 6).

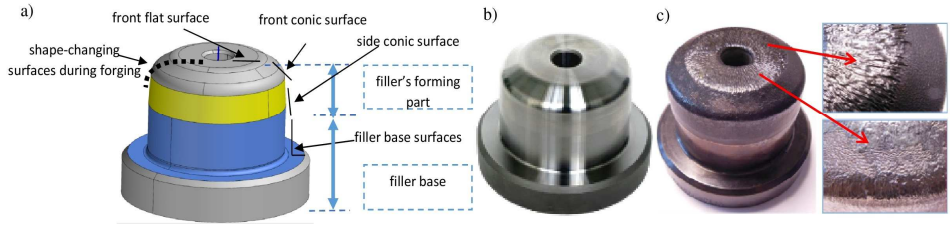

Fig. 6. A functional division of the punch surface with marked surfaces, which change their shape during forging (a); new tool (b); worn punch with with enlarged areas of wear (c).

The analysed filler can be divided into a shaping part and a base (Fig. 6). The base of the filler is responsible for the basing of the surfaces shaping the forging in respect of the other part of the assembled forging tool. The surfaces shaping the forging belonging to the moulding part of the filler (flat front, conical front, conical side) change the geometrical shape together with the number of produced forgings and are responsible for assuring the geometrical characteristics of a final product of the forging process in the second operation. The authors of the study [5] performed a thorough analysis of various variants of reference surfaces, which made it possible to obtain the highest convergence of wear results based on the analysis of the tool scans.

3.1.2. Comparison of tools after work (forging) with either image of new (unworn) tool or its CAD model

In the case of tools with "flat" and convex working surfaces, such as the ones shown in Figs. 4–6, the wear analysis is performed by way of comparing two scan images obtained for a new tool (before work) and a worn tool (after work). In the case when the forging tools, *e.g.* extrusion dies, have deep impressions, it is impossible to measure them before their work. Then, a worn tool after its work is cut into two parts, and its scan image is compared with its CAD model. An example of such analysis can be a die used in the second operation of a multi-operation process of forging a constant velocity joint boot (Fig. 7).

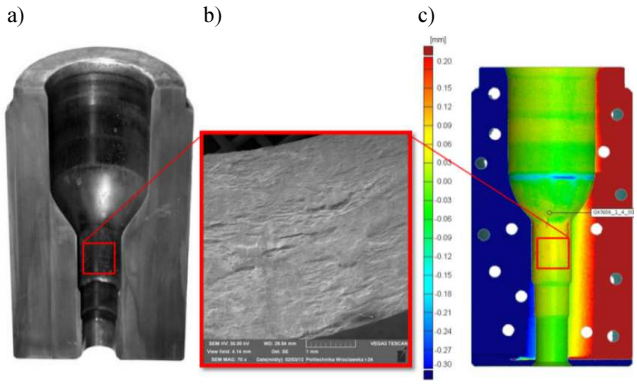

Fig. 7. Wear analysis of a tool with a deep working cavity – a die used to forge a CVJB: a view of the working surfaces of the cut die after work (a); a micro-area with adhesive wear (b); a scan image of the working surface (c) [8].

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

3.2. Advanced wear analysis

The use of a mobile measuring arm equipped with a scanner can be much more advanced, *e.g.* in the evaluation of changes in the working surfaces of a forging tool as well as progress of its wear. The authors performed measurements of wear of a selected forging tool after an increasing number of produced forgings, by comparing the obtained scan images of many worn tools with the scan image of a new tool (Fig. 8). The presented results of superimposing the images of the worn tools (after an increasing number of produced forgings) point to a progressive wear. In the initial period, for inserts after a small number of forgings, up to 1850 items, practically no material loss is visible. In turn, from 2500 forgings up, we can observe a clear wear in the central part and an increasing wear on the tool bridge. For an insert after producing 12500 forgings, the loss at the front equals to over 1.8 mm, and on the bridge – to about 1.6 mm. It can be seen that, for most of the tools, the wear at the insert front is clearly asymmetrical, while a more uniform wear can be observed on the bridge.

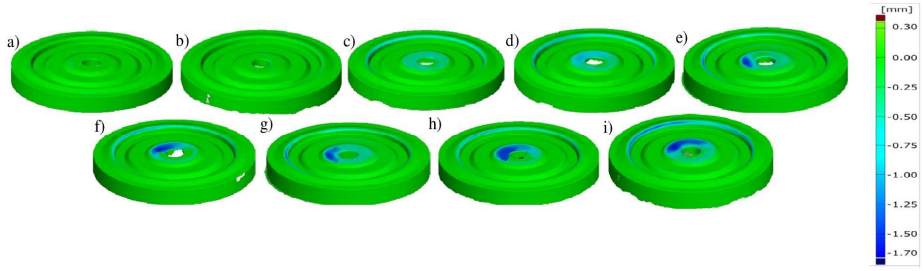

Fig. 8. Results of tool scanning after different numbers of produced forgings: 550 (a); 1850 (b); 2500 (c); 4300 (d); 5000 (e); 6900 (f); 9500 (g); 11000 (h); 12500 (i).

The analysis presented in Fig. 8 can turn out insufficient, and so, based on the collected scan images of a tool that produced an increasing number of forgings, one can elaborate the wear characteristic for this tool in a function of the increasing number of forgings, from 0 to 12500 items. Based on the presented diagram (Fig. 9), resembling the classic (Lorenz) wear curve, one can observe interesting relations and differentiate between scopes (periods) of wear. The presented analysis refers to the volumetric loss from all the working surfaces of selected tools, which can cause certain differences between particular scan images from Fig. 8.

We can see in the diagram (Fig. 9) that the material loss for a selected tool-die insert, based on the scan image analysis, increases very rapidly at the beginning of the forging process to up to about 2000 forgings (period I). This is connected with the adjustment of the whole system, in which we observe a transformation of the initial state of surface layers of co-acting elements of the tool with the forging into the optimal state. However, based on the mathematical analysis of the diagram itself, it can be stated that period I extends to up to 5000 forgings. In turn, on the basis of our own studies, we established that, from about 2000 forgings up, in most die forging processes, almost all types of degradation mechanisms begin to occur (abrasive wear, thermal and thermo-mechanical wear, plastic deformation, fatigue cracking). The intensity of these mechanisms depends mainly on: pressures, a contact time, temperature, a path of friction, the number of forgings, *etc*., and these directly translate to a given area in the tool [8, 10, 13].

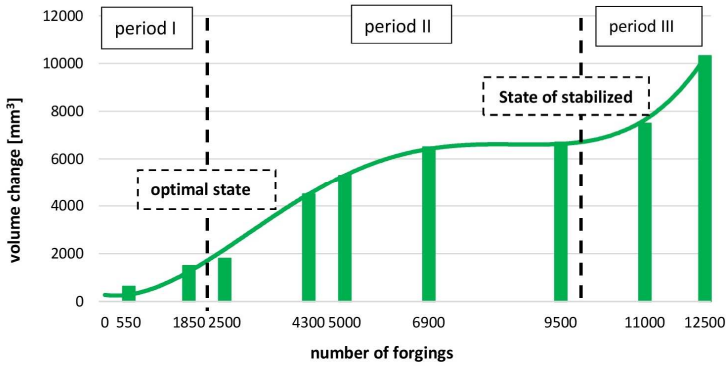

Fig. 9. A material loss (volume changes) from dies in a function of the number of forgings.

The performed research showed that, if a working surface is divided into smaller areas, characteristic for a given mechanism, the interpretation of particular wear diagrams will be identical with the mathematical one.

After reaching the optimal state, that is for over 3000 items, a so-called state of normal operation begins (period II), characterized by an approximately stabilized level of intensity of the mentioned degradation phenomena, which, in the analysed case, reaches the number of about 9500 items. A change in the volume for this scope of forgings equals from 3000 to 6200 mm^3, whereas, for the number of 9500 items up to the end of the tool's operation (over 12000 items), the volume change equals to as much as 4000 mm^3. On this basis we can conclude that the state of stabilized wear can be assumed to be for 3000–9500 forgings, which we can establish as the beginning of wear period III. This state, for the analysed tool, occurs to up to the maximal wear, that is to 12000 items, and ends at the moment when the acceptable change in the tool shape is exceeded, causing its removal from further production. A similar situation takes place for the classic Lorenz curve, which, close to the end of the normal operation period, usually transforms into the state of accelerated wear. It should be emphasized that the shape of the wear curve can differ in the case of other tools, which has been confirmed by the authors' studies, presented in [11].

3.3. Elaboration of 3D reverse scanning method

The following stage of the development of methods of tool wear analysis will be the construction of wear characteristics without the necessity of intervention into the executed forging process.

In this case, the 3D scanning method was used for an indirect control of the quality and geometry changes of forging tools (without the necessity of their disassembly) by way of a direct measurement of the geometrical changes of periodically collected forgings. On this basis, precise wear characteristics are constructed, whose result is comparable with the curve obtained based on the tool scans. The essence of the elaborated technology of die wear evaluation is the use of the changes in the forging shape which occur as a result of the die's wear during the forging process. To that end, the authors used the observed similarity (reflections) of the tool's working surface on a selected forging surface, in which the material loss of the tool is equal to the material growth on the forging. Fig. 10 shows the surface of a die before and after operation, together with the corresponding surfaces of the produced forgings. Fig. 11 shows the measured values of loss on the tool and their corresponding material allowances present on the analysed forging. The presented idea of reverse scanning uses the reflection of changes in the tool on the selected forging surface.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

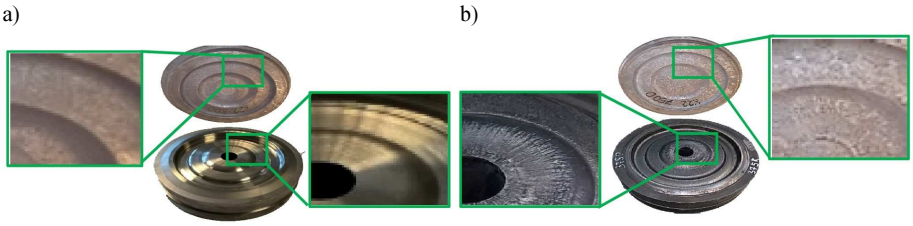

Fig. 10. An example of die insert: new – before work, together with the forging from the beginning of operation (500 items) (a); worn out – after producing 7 500 forgings, together with the forging from the end of the tool's operation (b).

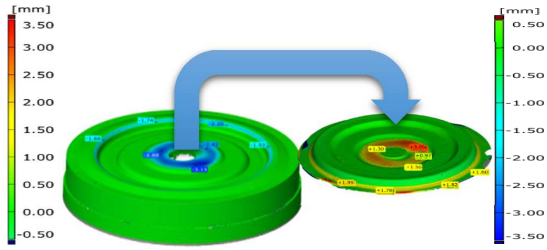

Fig. 11. A diagram of the proposed method of 3D scanning. Comparison of the scan images of a die and the last forging in the form of a shape change of a selected surface.

The elaborated method consists in applying a scanner to measure the proceeding wear of a selected forging tool (in the form of its material loss) based on the shape changes of periodically collected forgings (in the form of material growth of the forging) (Fig. 11).

3.3.1. Surface selection for measurement data equalization during measurements of forgings

In order to analyse the wear based on the periodically collected forgings, it is necessary to measure successive forgings on the measuring bench shown in Fig. 2b, or directly on the production line.

The results of analysis of typical material loss at the end of the tool's operation make it possible to determine the surfaces where the wear is scant in the forging's impression. The determination of these surfaces on a die insert makes it possible to point to the surfaces for measurement data equalization in order to perform the dimensional analysis of the forgings. The effect of proper determination of such a surface was discussed by the authors in the studies [5]. Fig. 12 shows selected surfaces where the tool wear is scant. They are necessary for a proper comparison of the measurement data in the forgings' analysis performed in the following stage of the elaborated measurement technology.

In the case of analysing the wear of the forgings, the scanned data are compared by means of the POLYWORKS software, with the use of the best-fit data equalization. In this process, as the reference surfaces, surfaces of the 100th forging were selected (Fig. 13a), which are formed in the die forging process on surfaces of the die with minor wear. In the considered cases, for the forging, it is the geometry shown in Fig. 13b. Fig. 13d shows a result of comparison of the scan images of the 12 000th (Fig. 13c) and the 100th forgings (Fig. 13a), using a reference for the best-fit equalization shown in Fig. 13b.

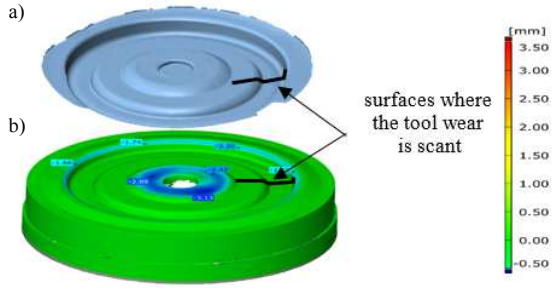

Fig. 12. An example of analysis results in the form of a map of deviations for a die insert with marked surfaces selected for the basing of forgings on the die (a); on the forging (b).

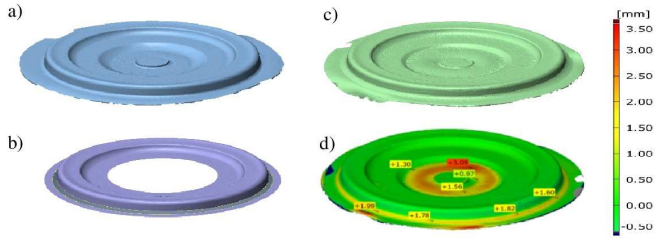

Fig. 13. A scan image of the 100th forging (a); a reference for best-fit equalization, for periodically collected forgings in the form of elements selected from the scan image of the 100th forging (b); a scan image of the 12 000th forging (c); a result of comparison of the 12 000th and the 100th forgings (d).

The result of comparison (Fig. 13d) is a coloured map of deviations distributed on the surface of the forging's scan image, showing the deviation value of error of the selected surface from the nominal dimension, which was the scan image of the 100th forging.

The result of measurements in the 3D scanning technology is a cloud of points. Next, on its basis, using the program, the polygonal surface was calculated, which consisted of triangle elements, reflecting the geometry of the measured object. In order to reconstruct the die wear course, the forgings, selected from the total number of 12 500 items for the selected die, underwent scanning (every 100 and 1000 items).

Figure14 shows scan images of the forgings (every 1000 item) obtained for a die. The results are presented in the form of shape changes of the selected surface in reference to the scan image of the 100th forging, which were obtained according to the measurement technology described above.

The presented results of the die wear analysis in Fig. 14 for an increasing number of forgings point to a proceeding wear of the tool, owing to the use of the reflection of tool image on the surfaces of successive forgings and their comparison with the "unworn" 100th forging.

The wear is located in the central part, in the vicinity of the pusher opening in the front area of the forging and, in the initial stage of the forging process, it is irregular. At the end of the die's durability period, one can see radial grooves on the deep ring (Fig. 14). In the scan images, wear in the area right in front of the bridge can be noticed (vicinity of the flash), in the form of an asymmetrical ring of wear.

The presented results in the form of a deviation in the shape of periodically collected forgings make it possible only to perform a simplified analysis. The latter enables the determination of the die areas where the wear occurs, as well as the areas of the maximal material loss. Such a reconstruction of the wear course makes it possible to perform an analysis at a time interval corresponding with the frequency of collection of forgings.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

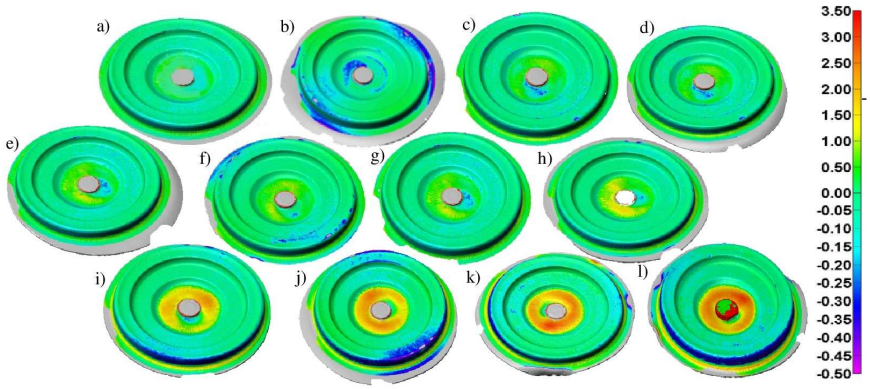

Fig. 14 Comparison of scan images of forgings made in a die, in the form of a shape change in a selected surface, referred to the 100th forging, after: 1000 (a); 2000 (b); 3000 (c); 4000 (d); 5000 (e); 6000 (f); 7000 (g); 8000 (h); 9000 (i); 10000 (j); 11000 (k); 12000 items (l).

The results of measurement by way of scanning of the last forging were compiled with the results of measurement of the die at the end of the forging process (Fig. 15).

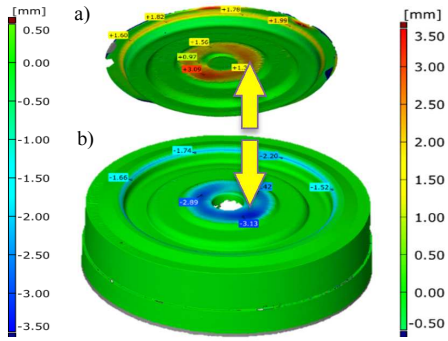

Fig. 15. Compilation of 3D scan results ofthe last forging (a); the die at the end of the forging process (b).

The presented results of the wear analysis calculated based on the analysis of the forging are very similar to the results of a typical analysis of the die wear performed at the end of the forging process. And so, it can be assumed that with the use of the tool change reflection, during the forging process, on the periodically collected forgings, the obtained results are convergent and make it possible to perform an analysis of wear.

3.3.2. Determination of Lorenz curve based on 3D reverse scanning

An expansion of the method of forging tool wear analysis based on a forging measurement is an analysis which uses the volumetric change occurring during the process of die wear. Such an approach enables a global and thorough description of the phenomenon of material loss during the forging process by way of measuring the systematically collected samples.

In order to determine a diagram describing the dependence of the volumetric wear on the number of produced forgings during the forging process, it is necessary to calculate the volume change in the forging areas selected at an earlier stage, marked with circles in Fig. 16.

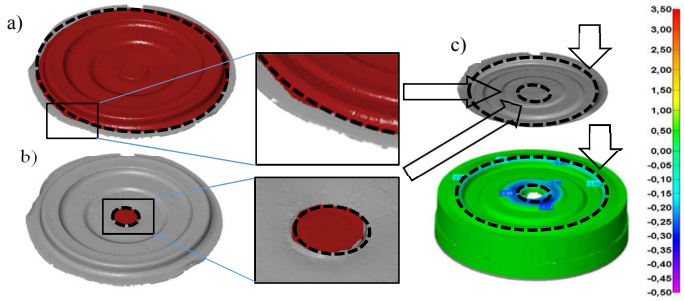

Fig. 16. An example of volume measurements in the areas determined by circles the total volume for the area (a); the core volume for the area (b); a schematic diagram of the volume measurement (c).

During the measurement, an algorithm is applied, consisting in measuring the volume in the areas determined by circles with the use of the POLYWORKS program. The selected software makes it possible to fill the volume between two previously equalized surfaces constructed from scanned triangles generated by means of the Real-Time Quality Meshing technology during the scanning.

The above assumption made it possible to calculate the volumes (Figs. 16a and b), which, when one was subtracted from the other, enabled to determine the desired values of wear in the considered ring-shaped areas of the forging (Fig. 16c).

Based on the volumes determined in the analysis for each periodically collected forging, it is possible to construct a diagram showing the material loss (volume changes) of a worn tool. In this way, it is possible to determine a tool wear curve (Lorenz curve) based on the forgings. Such an approach is a much more practical solution, as it does require neither disruption of the production process nor disassembly of a selected tool after a specific number of forgings.

The reverse scanning method has already been implemented into the industrial forging process of lid forging in the wear analysis of a tool shown in Fig. 6. Details on the implementation of this method are described in the paper [11].

A confirmation of effectiveness of the elaborated method of wear analysis with the use of reverse scanning is, of course, a comparison of both wear curves, determined based on the results of scanning of the worn tools and the periodically collected forgings.

3.3.3. Comparison of Lorenz curve determined based on scanned tools and 3D reverse scanning

Figure 17 shows a comparison of the Lorenz curve determined based on the scanned tools after an increasing operation time (Fig. 8), as well as that based on the method of 3D reverse scanning by way of measurement of the systematically collected forgings (their scan images are shown in Fig. 14).

The comparison, shown in Fig. 17, of both methods of determining the relations describing the tool wear during the forging process (determining their durability), points to a significant convergence. The highest divergences can be observed at the very beginning, that is from 0 to 2500 forgings, as well as in the scope from 4500 to 9500 forgings. Analysis of the curves in the Fig. 17 is similar to those shown in Fig. 9. That is, the stabilized state is obtained earlier than it would look on the basis of mathematical curve analysis.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

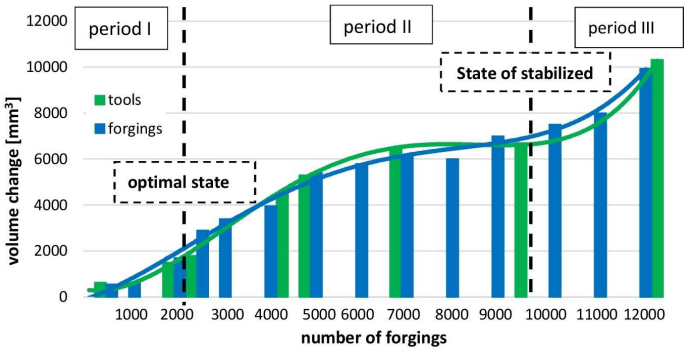

Fig. 17. Comparison of wear curves based on tool (green) and forging (red) scan images in a function of the number of forgings.

The differences in the initial scope probably result from the stabilization of the process (the whole system), that is from the adjustment of a proper temperature of tool operation and optimal conditions of lubrication and cooling (tribological conditions). In turn, the differences in the later period can be explained by the studies performed by the authors, which demonstrated that, for this process, from about 4000–5000 forgings up, one can observe intensification of the destructive mechanisms, which is connected with the detachment of larger particles of the nitride layer and the tool material from the most loaded areas.

Another cause of the small divergences between both curves can be the fact that the forgings for the determination of the Lorenz curve were collected (every 1000 items) from one forging process, for which an average durability is equal to 12 000 forgings, while the tools selected for the determination of wear were collected from a few similar processes, yet after an increasing number of produced forgings. This was dictated by the idea of maintaining similar technological conditions (elimination of the process of cooling the tool for a scan analysis, followed by its heating for the further production process). Also, each tool, after the scan analysis, did not come back to operation, but was cut into samples for further tests, such as: microstructural tests, SEM (Scanning Electron Microscope), micro-hardness measurements, *etc.* Other, less important, causes of the minimal divergences can be a measuring accuracy of the scanner (based on a volumetric performance test according to the standard B89.4.22, its precision is equal to 0.058 mm), as well as the oxidation and scale coating of the measured forgings which were cleaned before the measurement. as Another one can be the errors resulting from the calculation algorithm in the volumetric analysis.

In the case of the presented 3D scanning method (Figs. 9 and 17), the forgings were measured in laboratory conditions with no vibration and constant temperature. The only limitation is the removal of scale from the surface of forgings. The dismantled tools are also scanned in a similar way to the measurement of forgings. So, in this case it is difficult to talk about a significant influence of temperature and vibration (technological break), although the latter are actually present in the production hall. Also, an impact of vibration on the measurement results was not observed during the scanning of tools on the press.

Considering the above, the presented comparison confirms that the determination of wear based on the scanning of the forgings periodically collected during the technological process (without the necessity of their disassembly) is an effective and economically justified method. It should also be emphasized that the determination of wear based on tool scans is an impractical method, generating additional cost and often causing difficulties in maintaining the continuity of production, as well as its disruption and changes in the technological and tribological conditions.

3.4. Expanded 3D reverse scanning method

The presented method of analysing the wear of forging tools can be expanded by a division into additional areas. Such an expansion can be dictated by the necessity of considering different degradation mechanisms. It will also enable to eliminate errors during the total volumetric analysis, occurring in the areas of cracks or excessive local tool wear. The standard use of diagrams of the volume changes occurring during the process of wear in the analysed areas provides additional information. Such an approach enables a comprehensive description of the phenomenon of material loss during the forging process. However, in certain cases, the analysis of the total material loss of a given tool does not provide a full image (Fig. 18). In such special cases, the comparative analyses of a few such tools can contain errors resulting from *e.g.* tool cracking or premature wear of one of the areas, which will distort the calculation results of the material loss volume.

Figure 18a shows an image of a die insert with a division into two selected characteristic areas (A and B), for which, based on the preliminary analysis, the occurrence of different degradation mechanisms was established. Additionally, the performed 3D reverse scanning analysis showed a much faster wear of the forging tool than in area A. In this area, a cross is made in the new tool which plays the role of a marker. As it can be observed in Fig. 19b, the wear in areas A and B is almost identical for up to 2500 forgings. Above this number, the wear intensively increases at the front of the insert (area A), in respect of the wear on the bridge (area B). That is why the performed analysis of the total wear of the insert can be loaded with error resulting from a different intensity of tool wear in different areas of the tool.

In order to perform a precise analysis and create a diagram (Fig. 18) describing the volumetric wear in areas A and B depending on the number of produced forgings, during the forging process, it is necessary to calculate the volume change in the areas selected at the earlier stages. The elaborated measurement technology employs an algorithm consisting in measuring the volume in the areas marked by circles with the use of the POLYWORKS software.

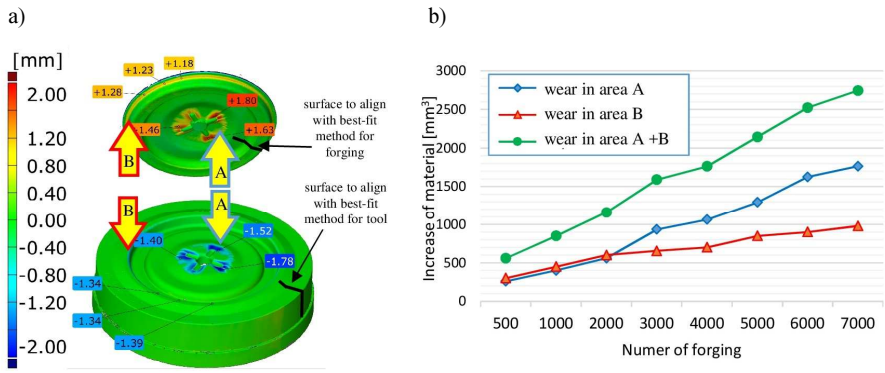

Fig. 18. A method of 3D reverse scanning – division into selected areas (a); wear based on periodically collected forgings (b).

The selected program makes it possible to fill the volume between the two previously levelled surfaces constructed from the scanned triangles generated with the use of the Real-Time Quality Meshing method. To that end, four circles were assumed in the program, two for each of two areas, A and B.

This assumption enabled the calculation of the volumes, which subtracted in pairs, one from the other, made it possible to determine the two desired values of wear in areas A (Figs. 19a and 19b) and B (Figs. 19c and 19d), as well as the total value, that is the sum of A and B.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

Based on the calculated volumes for A and B for each analysed, periodically collected, forging, it is possible to construct a diagram presenting the material loss (volume changes), calculated based on the volume changes between the surfaces of the forgings produced on the analysed die, which is shown in Fig. 19b.

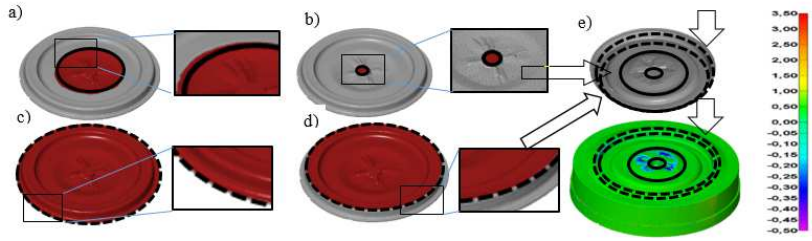

Fig. 19. An example of determining the volume in selected areas: the core volume for area A (a) the total volume for area A (b); the core volume for area B (c); the total volume for area B (d); a schematic diagram of the volume measurement (e).

3.4.1. Pros and cons

The industrial application of the reverse scanning method showed advantages and disadvantages of the elaborated technology. The basic advantage of the 3D reverse scanning method is its practicality in being used in the hard conditions of die forges. For example, Fig. 20 shows the measurement of a relatively large forging of a scraper weighing over 42 kg and a heavy lower die used during forging with a pneumatic hammer MPM 16000 with an impact energy of 171,62 kJ (a mass of the component elements is equal to 5285 kg without the die; a mass of the tool is over 900 kg). In such a case, a fast analysis of the progress of wear of heavy forging dies, based on the measurement of periodically collected forgings, is irreplaceable.

In turn, the main disadvantages of this method are the error resulting from the hindered measurement of a warm forging, which temperature, in extreme cases, is about 150–250ºC and the fact of taking into account the thermal expansion (for the QS1920 steel and this temperature range it is about $1.45 \cdot 10^{-5}$/K). Also, the scale on the surface of forgings, whose thickness, in many cases, is high enough (even over 0.5 mm) to influence the results of both the measurement and the analysis of tool wear based on the forgings. For the 3D scanning method, a verifying measurement of the last forgings – that is the measurement of the tool – is performed during a maintenance shutdown. The working temperature of tools is equal to about 200–250ºC on the surface.

Fig. 20. Measurements of a ”hot” forging of a scraper (the indirect reverse scanning method) and a forging die on a hammer, during a technological break (in order to verify the reverse method).

During the shutdown, the tool is cooled down to about 80ºC and cleaned before the measurement. The scanning procedure itself is executed a dozen or so minutes after the press has been put to a stop.

3.5. Directions of development

The use of scanning methods for measurements in the forging industry is at present very extensive. Of course, this development can be considered in two ways. One direction of the development of scanning methods are the measurements of the forgings for the control of their geometry and quality. The other direction is the application of scanning to the analysis of forging tools.

A prospective direction of the development of the reverse scanning method can be the analysis of the tool wear in more than two areas, because, as we know, the shape of the tool will determine the occurrence of various degradation mechanisms in different areas. Fig. 21 shows the areas in the impression of a die, where shape-determined degradation mechanisms occur. As one can notice, abrasive wear will be dominating in the areas where the deformed material intensively moves, filling the die impression. In the areas of stress concentration, that is the ones with small internal radii in the impression, sharp edges *etc*., we will observe mechanical and thermo-mechanical fatigue causing brittle cracking. The areas of a long contact of hot forging material will be characterized by the occurrence of plastic deformations caused by local material tempering as well as thermal fatigue, which will also be present together with abrasive wear in the areas where the hot material flows, thus intensifying the degradation process of the substrate in these areas.

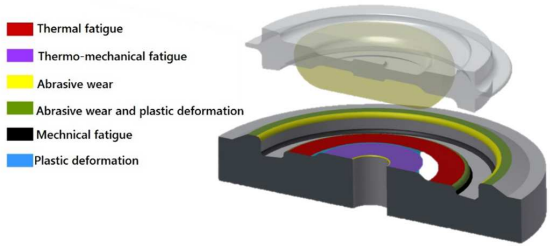

Fig. 21. A diagram illustrating the areas of dominating degradation mechanisms in the impression of a forging die.

Such a detailed analysis of tool wear, in a few or a dozen areas, especially based on the scanning of forgings, provides very valuable information (on the quantitative material loss), which can be used *e.g*. in the construction of a data base in expert systems, which enable the prediction of the forging tool durability [10].

4. Conclusions

At present, in the process of production of forgings, in die forges, different devices and measurement methods are used. The latter include methods applying the universal, classical measurement equipment, ensuring lower measurement accuracy for the control of the key geometrical characteristics in a hot forging, through more complicated measurement techniques using the universal measurement equipment, which, in combination with specially designed tests, makes it possible to fully control the tool quality and properties, of a non-complicated geometry, to measurement methods based on Coordinate Measuring Technique as well as scanning techniques for both forgings and forging tools of a complicated geometrical specification. Also,

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

new trends are being created, which are mainly connected with the possibility of applying portable measurement systems, such as: optical scanners or linear scanners mounted on portable measuring arms. This technology enables effective control of the quality of medium- and large-sized forgings, as well as forging tools of very large sizes, whose measurement takes place directly on the production line.

The methods of 3D scanning and forging tool wear analysis discussed in the study make it possible to perform simple as well as complex and extensive analyses of the wear of forging tools and instrumentation.

The presented results of the studies performed with the use of a measuring arm together with an integrated laser scanner for the analysis of wear, enabled the elaboration of a 3D reverse scanning method based on measurements of the shape changes in the successive forgings (directly on the production line). The performed studies showed the validity of using new, non-contact measurement technologies for a direct and indirect analysis of tool quality and shape changes (without disassembling the instrumentation from the forging aggregate). Owing to the implementation of non-contact measurement methods, the analysis of tool wear has become possible to be performed directly on the production line.

Nevertheless, the presented approach to measurements, assuming the possibility of using more accurate scanning techniques, can be used to analyse tools in plastic moulding, foundry, as well as in the food industry [22, 23, 31].

The analysis of the volume increase of the successive forgings based on the measurements makes it possible to precisely determine the material loss of a forging tool in the successive stages of its performance. This is proved by the full correlation between the results of measurements of the volume changes of an increasing number of produced forgings and those of the tool, which is suggested by the verification of the wear (Lorenz) curves performed in the study.

The innovative approach to evaluation of the current state of a forging tool performed by the authors makes it possible to make decisions about a prolongation or shortening of its operation time, based on the actual (current) wear, rather than based on the fixed durability data (a maximal number of produced forgings). This enables an optimal use of a given tool, preserving the highest quality of produced forgings.

The demonstrated advantages and disadvantages of the proposed approach to analysis of tool wear with the use of 3D scanning certainly make it possible to prolong the operation time of forging instrumentation and significantly lower the production costs.

Acknowledgment

The research has been financed by the National Centre of Research and Development (NCBiR); project no. POIG.01.03.01-02-063/12.

References

[1] Bewoor, N.A., Kulkarni, V. (2009). *Metrology & Measurement Paperback*. New Dehli: Tata McGraw-Hill Education Private Limited.

[2] Bills, P.J., Racasan, R., *et al*. (2012). Volumetric wear assessment of retrieved metal-on-metal hip prostheses and the impact of measurement uncertainty. *Wear*, 274–275, 212–219.

[3] Carmignato, S., Spinelli, M., Affatato, S., Savio, E. (2011). Uncertainty evaluation of volumetric wear assessment from coordinate measurements of ceramic hip joint prostheses. *Wear*, 270 (9–10), 584–590.

[4] Cuesta, E., *et al*. (2015). A statistical approach to prediction of the cmm drift behaviour using a calibrated mechanical artefact. *Metrol. Meas. Syst.*, 12(3), 417–428.

[5] Dworzak, Ł., Hawryluk, M., Kaszuba, M., Ziemba J. (2016). Analysis of the approximation of data obtained from scanning of a forging instrument measuring arm under production conditions. *IV International Scientific Technical*

Conference MANUFACTURING 2014 8–10 Dec. 2014 selected conference proc./Poznan University of Technology, Poznań, Poland, 23–34.

[6] Gronostajski, Z. *et al.* (2011). Application of the scanning laser system for the wear estimation of forging tools. *Computer Methods in Materials Science*, 11(2), 425–431.

[7] Gronostajski, Z., *et al.* (2015). Improving durability of hot forging tools by applying hybrid layers. *Metallurgy*, 54(4), 687–690.

[8] Gronostajski, Z., Kaszuba, M., Hawryluk, M., Zwierzchowski, M. (2014). A review of the degradation mechanisms of the hot forging tools. *Archives of Civil and Mechanical Engineering*, 14(4), 528–539.

[9] Gronostajski, Z., Hawryluk, M., Jakubik, J., Kaszuba, M., Misun, G., Sadowski, P. (2015). Solution examples of selected issues related to die forging. *Archives of Metallurgy and Materials*, 60(4), 2767–2775.

[10] Gronostajski, Z., Hawryluk, M., *et al.* (2016). The expert system supporting the assessment of the durability of forging tools. *International Journal of Advanced Manufacturing Technology*, 82, 1973–1991.

[11] Gronostajski, Z., Hawryluk, M., Kaszuba, M., Ziemba, J. (2016). Application of a measuring arm with an integrated laser scanner in the analysis of the shape changes of forging instrumentation during production. *Eksploatacja i Niezawodnosc – Maintenance and Reliability*, 18(2), 194–200.

[12] Gutiérrez, R., Ramírez, M., Olmeda, E., Díaz, V. (2015). An uncertainty model of approximating the analytical solution to the real case in the field of stress prediction. *Metrol. Meas. Syst.*, 22(3), 429−442.

[13] Hawryluk, M., Zwierzchowski, M., Marciniak, M., Sadowski, P. (2017). Phenomena and degradation mechanisms in the surface layer of die inserts used in the hot forging processes. *Engineering Failure Analysis*, 79, 313–329.

[14] Hawryluk, M., Jakubik, J. (2016). Analysis of forging defects for selected industrial die forging processes. *Engineering Failure Analysis*, 59, 396–409.

[15] ISO GPS 10360-4:2000 Geometrical Product Specifications (GPS) – Acceptance and Reverification Tests for Coordinate Measuring Machines (CMM) – Part 4: CMMs used in Scanning Measuring Mode. Norm.

[16] Kuş, A. (2009). Implementation of 3d optical scanning technology for automotive applications. *Sensors*, 9, 1967–1979.

[17] Langton, D.J., *et al.* (2014). Practical considerations for volumetric wear analysis of explanted hip arthroplasties. *Bone & Joint Research*, 3, 60-68.

[18] Li, F.X., Longstaff, A., Fletcher, S., Myers, S. (Mar 2012). Integrated tactile and optical measuring systems in three dimensional metrology. *Computing and Engineering Researchers' Conference*, University of Huddersfield, 1–6.

[19] Lord, J.K., Langton, D.J., Nargol, A.V.F., Joyce, T.J. (2011). Volumetric wear assessment of failed metal-on-metal hip resurfacing prostheses. *Wear*, 272(1), 79–87.

[20] Lu, Z., McKellop, H.A. (2014). Accuracy of methods for calculating volumetric wear from coordinate measuring machine data of retrieved metal-on-metal hip joint implants. *Proc. of the Institution of Mechanical Engineers*, part H, 228(3), 237–49.

[21] Machácek, P., Tomícek, J. (2010). Application of laser scanning in reverse engineering and prototype manufacturing. *WTP*, 1(21), 35–44.

[22] Majchrowski, R., Grzelka, M., Wieczorowski, M., Sadowski, Ł., Gapiński, B. (2015). Large area concrete surface topography measurements using optical 3d scanner. *Metrol. Meas. Syst.*, 22(5), 565–576.

[23] Mathia, T., Pawlus, P., Wieczorowski, M. (2011). Recent trends in surface metrology. *Wear*, 271(3–4), 494–508.

[24] Measurement of a forging die for tooling corrections. APPLICATION REPORT: AICON_3D_SYSTEM. *http://www.aicon3d.com/fileadmin/user_upload/produkte/en/breuckmann_Scanner/01_PDF_IuT/Forging_die_measurement_Web.pdf. (June 2017).*

[25] Peng, X., Wang, L., *et al.* (2014). Optimized adaptive control for evaluation of artificial hip joint volumetric wear by coordinate measuring. *Hsi-An Chiao Tung Ta Hsueh/Journal of Xi'an Jiaotong University*, 48(8), 128–135.

[26] Raghavendra, N.V., Krishnamurthy, L. (2013). *Engineering Metrology and Measurements Paperback*. Oxford: Oxford University Press.

[27] Ratajczyk, E., Woźniak, A. (2016). *Coordinate measuring systems*. Warszaw: Oficyna Wydawnicza Politechniki Warszawskiej.

Metrol. Meas. Syst., Vol. 24 (2017), No. 4, pp. 739–757.

[28] Salah Hame, R.A. (2008). Influence of fitting algorithm and scanning speed on roundness error for 50 mm standard ring measurement using CMM. *Metrol. Meas. Syst.*, 15(1), 33–53.

[29] Sładek, J., Sokal, G., Kmita, A., Ostrowska, K. (2007). Calibration of coordinate measuring arms. *Acta Mechanica et Automatica*, 1(2), 53–58.

[30] Weckenmann, A., Weickmann, J. (2006). Optical inspection of formed sheet metal parts applying fringe projection systems and virtual fixation. *Metrol. Meas. Syst.*, 13(4), 321–334.

[31] Wieczorowski, M., Ruciński, M., Koteras, R. (2010). Application of optical scanning for measurements of castings and cores. *Archives of Foundry Engineering*, 10, 265−268.

[32] Yu-cun, Z.G., Jun-xia, H., Xian-bin, F., Fu-li, Z. (2014). Measurement and control technology of the size for large hot forgings. *Measurement*, 49, 52–59.

[33] Yu-cun, Z., Bin, W., Xian-bin, F. (2014). An unsteady temperature field measurement method for large hot cylindrical shell forging based on infrared spectrum. *Measurement*, 58, 12–20.

[34] Zhengchun, D., Zhaoyong, W., Jianguo, Y. (2016). 3D measuring and segmentation method for hot heavy forging. *Measurement*, 85, 43–53.

[35] Zhenyuan, J., Bangguo, W., Wei, L., Yuwen, S. (2010). An improved image acquiring method for machine vision measurement of hot formed parts. *Journal of Materials Processing Technology*, 210, 267–271.

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.edu.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