RENSIT:
RadioElectronics. NanoSystems. Information Technologies

Journal "Radioelectronics. Nanosystems. Information Technologies" (abbr. RENSIT)
publishes original articles, reviews and brief reports, not previously published, on topical problems in **radioelectronics (including biomedical) and fundamentals of information, nano- and biotechnologies and adjacent areas of physics and mathematics.**

Designed for **researchers, graduate students, physics students of senior courses and teachers**.
It turns out **2 times a year** (that includes 2 issues)

Authors of journal are academicians, corresponding members and foreign members of Russian Academy of Natural Sciences (RANS) and their colleagues,
as well as other russian and foreign authors on presentation of their manuscripts by the members of RANS, which can be obtained by authors before sending articles to editors.
And also after its receiving - on recommendation of a member of editorial board of journal, or another member of Academy of Natural Sciences, that gave her opinion on article at request of editor.
The editors will accept articles in both **Russian and English** languages.

Articles are internally peer reviewed (**double-blind peer review**) by members of the Editorial Board.
Some articles undergo external review, if necessary.

Journal RENSIT is included in the **DB SCOPUS, EBSCO Publishing,** in the international abstracts database - **Ulrich's International Periodicals Directory**,(USA, New York, http://www.ulrichsweb.com), in the **AJ and DB VINITI RAS** (http://www.viniti.ru), and DB **Russian Science Citation Index (RSCI** http://elibrary.ru/project_risc.asp).

Full-text content is posted in the DB of the **Russian Scientific Electronic Library**
- information resource on the Internet http://elibrary.ru and is available for registered users.
And also - in Open Access **CyberLeninka NEB** of Russian Federation http://cyberleninka.ru.

On journal's website http://www.rensit.ru posted metadata publications
and **RENSIT: Radioelectronics. Nanosystems. Infomation Technologies - english version** (cover-to-cover translation) of journal, which is a party to **CrossRef.**

The founder - the **Russian Academy of Natural Sciences**
Publisher - Publishing Center of the Russian Academy of Natural Sciences
Publisher Address: 29/16, Sivtsev Vrazhek lane, Moscow 119002, Russian Federation

CONTENTS

RUSSIAN ACADEMY OF NATURAL SCIENCES

DEPARTMENT OF RADIOELECTRONICS, NANOPHYSICS AND INFORMATION TECHNOLOGIES PROBLEMS

RENSIT:

RADIOELECTRONICS. NANOSYSTEMS. INFORMATION TECHNOLOGIES.

2020, VOL. 12, № 1

FOUNDED IN **2009**
3 ISSUES PER YEAR
MOSCOW

ISSN 2414-1267

The journal on-line is registered by the Ministry of Telecom and Mass Cjmmunications of the Russian Federation. Certificate El. no. FS77-60275 on 19.12.2014

EDITORIAL BOARD ADDRESS
218-219 of., 7 b., 11, Mokhovaya str., 125009 MOSCOW, RUSSIAN FEDERATION,
TEL. +7 495 629 3368
FAX +7 495 629 3678 FOR GRACHEV

Track membranes and their replicas as high-frequency phase-contrast objects in X-ray optics

Alexander V. Mitrofanov

Lebedev Physical Institute of RAS, https://lebedev.ru/
Moscow 119991, Russian Federation
E-mail: mitrofanovav@lebedev.ru

Alexey V. Popov, Dmitry V. Prokopovich

Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation of RAS,https://izmiran.ru/
Troitsk, Moscow 108840, Russian Federation
E-mail: popov@izmiran.ru, dvprokopovich@gmail.com

Received December 10, 2019; peer reviewed April 07, 2020; accepted April 10, 2020

Abstract. **Possibility of using polymer track through membranes and membrane replicas as elements of X-ray optics for the visualization of microobjects with high spatial resolution is discussed. It is shown that samples prepared on the basis of track membranes and their replicas can be used in a wide X-ray range, including soft spectral regions (≥ 1 nm) as phase screens or model phase test objects for X-ray microscopy. Highly porous membranes, being diffuse weakly absorbing samples in the form of a single layer or stack of several films, influence coherent properties of the primary X-ray beam. Optical constants of the material of available porous membranes, their thickness, density and size allow one to vary optical characteristics of the phase screens in a wide frequency range, including visible region of spectrum and the X-ray band where many sources of synchrotron radiation work. The issues of wave field concentration by phase structures with narrow through channels, the effect of pore diameter on phase velocity in the channels and spreading (in transverse plane) of the phase pattern for membranes with extremely narrow pores, limiting the resolution of the phase screen used as X-ray test object, is studied in this work in detail. Numerical experiments have been performed by solving parabolic wave equation with tabular values of optical constants for membrane material.**

Keywords: **X-ray optics, phase-contrast X-ray microscopy, test nanoobjects, track membranes, parabolic equation method in X-ray optics**

PACS 41.50.+h, 41.60.Ap, 02.60.Cb

Acknowledgements: The authors thank P.Yu. Apel and A.A. Snigirev for useful discussions.
For citation: Alexander V. Mitrofanov, Alexey V. Popov, Dmitry V. Prokopovich. Track membranes and their replicas as high-frequency phase-contrast objects in X-ray optics. *RENSIT*, 2020, 12(2):173-190; DOI: 10.17725/rensit.2020.12.173.

Contents

1. INTRODUCTION

As it is known, the phase contrast method in optics was proposed by Dutch physicist F. Zernike in 1934 as a replacement of the "shadow method" being used for quality control of astronomical mirrors [2]. As soon as in 1935, he examined the application of the method of microscopic phase contrast imaging of refractive objects [3]. In 1953, F. Zernike received Nobel prize for the development of the method and creation of the first phase contrast microscope. The history of this discovery is presented in his Nobel lecture published, with minor changes, in 1955 in Science magazine [2]. In experimental optics the method of phase contrast relatively quickly became in demand and already before 1953 first reviews and theoretical papers appeared on Zernike's method [4, 5]. In the monograph by M. Franson [6], optical schemes and the principle of the phase contrast microscope designed for studying various phase objects in the visible spectral band (transparent micro-objects, reflecting layers with weakly pronounced surface relief, samples of biological tissues, etc.) are discussed. Methods and optical circuits for the phase-to-amplitude contrast conversion are being considered.

Recently phase contrast method got further development in X-ray microscopy [7]. Study of low-contrast biological objects at the cellular level, problem of testing microelectronic and nanotechnology products has become an impetus for development and improvement of optical schemes working on the principle of X-ray phase contrast microscope. Successes in this branch of experimental physics were achieved in recent years thanks, to a large extent, to the emergence of powerful X-ray sources and progress in focusing X-ray optics. Details are considered in the review [7]. Note that the estimated spatial resolution of X-ray microscopy lies between optical and electron microscopy. However, by the radiation dose imposed on the sample during the session, phase contrast microscopy is the most delicate and nondestructive technique compared with any research mode of the electronic microscopes.

The choice of the contrast type depends on the properties and size of the object itself, its preparation, technical characteristics of the equipment, radiation source, etc. Speaking on studying the internal structure of micro- and nano-objects, the test result at a wavelength is determined by the optical properties of the object – first of all, total refraction index of the sample, in which the X-ray beam propagates [8, 9]:

$$n(\lambda) = 1 - \delta + i\beta. \quad (1)$$

Here δ, β are the optical constants of the material, wavelength dependent. Usually they are small additions to unity in the expression for the refraction index (1). To the present time, dispersion of the refractive index of various substances and materials in a wide X-ray spectral region is well studied, and the quantities $\delta(\lambda)$ and $\beta(\lambda)$ can be found with the required accuracy from literary sources or well-known network resources [10]. In a classic X-ray band, at moderate photon energies $E \leq 10$ keV, radiation absorption in the sample is determined mainly by the photoelectric effect by the electrons transition from an atomic shell in the Coulomb field of the nucleus [11] (taking into account the field screening by the internal shells). Total absorption due to the photoelectric effect is described by the value β, and transmission coefficient of the sample at the wavelength λ, for a parallel beam, is given by the following expression:

$$T = \exp(-\mu L) = \exp(-4\pi\beta L/\lambda), \quad (2)$$

where μ is the linear absorption coefficient, L is the thickness of the sample. The real correction $\delta(\lambda)$ in (1) is positive, it arises as a result of the collective electron response of

each atom of the sample in the X-ray radiation wave field, like in the case of electrons in ionized plasma. Similar to the case of plasma in a high-frequency field, the real part of the refraction index (1) can be written in the form:

$$1 - \delta = 1 - (1/2)(\omega_c/\omega)^2, \quad (3)$$

where $\omega = 2\pi c/\lambda$ is the circular frequency, c is light velocity, ω_c – so-called critical frequency, of plasma, coinciding with the circular frequency of the electron gas free vibrations:

$$\omega_c = (4\pi N_e e^2/m)^{1/2} \quad (4)$$

(here, e and m are the charge and mass of the electron, N_e is electron density). Note that in contrast to the photoelectric effect all the electrons of the atom in a wide wavelength range (outside the absorption jumps) make approximately the same contribution to the real part of dielectric permittivity of the medium and therefore in the value of $\delta(\lambda)$, regardless of which shell contains the electrons. Using Eq. (4), the expression for the real part of deviation $\delta(\lambda)$ in Eq. (1) can be written in another form:

$$\delta = (N_e e^2/2\pi mc^2)\lambda^2, \quad (5)$$

or, since $N_e = ZN_{at}$, where Z is the atomic nucleus charge, N_{at} – atomic density, in the more familiar form [8-10] it reads:

$$\delta = (e^2/2\pi mc^2)\lambda^2 N_{at} Z \approx (e^2/2\pi mc^2)\lambda^2 N_{at} f_1, \quad (6)$$

where f_1 is the so-called atomic scattering function of the material (its real part), almost equal to the effective atomic number (outside the X-ray absorption jumps); here $e^2/\pi mc^2 = (2.81E^{-13})$sm – classical radius of the electron [8-10].

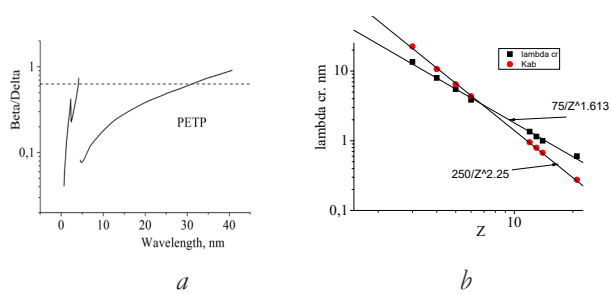
2. WHEN AN OBJECT IN X-RAY OPTICS CAN BE CONSIDERED AS PHASE SCREEN?

An ideal X-ray phase contrast object, L being maximum structural thickness of its elements, when a parallel X-ray beam with a wavelength λ passes through, provides, firstly, a noticeable (of order of unity) phase shift $2\pi\delta L/\lambda$ compared with the phase of the wave covering the same distance L in free space and, secondly, not very large absorption, so the beam attenuation in the sample does not prevent observing the differential phase pattern of the object in the transmission beam, i.e.:

$$2\pi\delta L/\lambda \approx 1,\ 4\pi\mu L/\lambda << 1. \quad (7)$$

In addition, for a given wavelength and experiment geometry, the measured phase shifts should not be too large that is an ideal for observation object must not be too thick- the first of the relations (7). Otherwise, if the reverse condition $2\pi\delta L/\lambda >> 1$ is satisfied, the measurement result will be ambiguous or utterly dependent on the sample tilt and jitter, beam angular characteristics, experiment layout, radiation monochromaticity degree etc. To a first approximation, the object can be considered rather phase than amplitude one when the inequality $\delta/\beta \geq 1$ begins to fulfil. If these two optical constants are of the same order the material phase shift over the characteristic damping length $\lambda/4\pi\beta$ equals 0.5 radian. This limiting length can be considered as a conditional boundary range when the object becomes phase one with decreasing X-ray wavelength.

As an example, in **Fig. 1***a* is depicted the ratio β/δ spectral dependence for polyethylene terephthalate (PETP) – polymer material of porous track membranes considered in this article (**Fig. 2**). The intersection of this curve with the dashed line marks the wavelength when, in terms of characteristic amplitude, the attenuation distance of a homogeneous



a *b*

Fig. 1. *(a) Optical constant ratio $\delta(\lambda)/\beta(\lambda)$ for PETP, as a function of wavelength. (b) Spectral dependence of threshold $\delta(\lambda)/\beta(\lambda) = 1$ for light elements with different* [10].

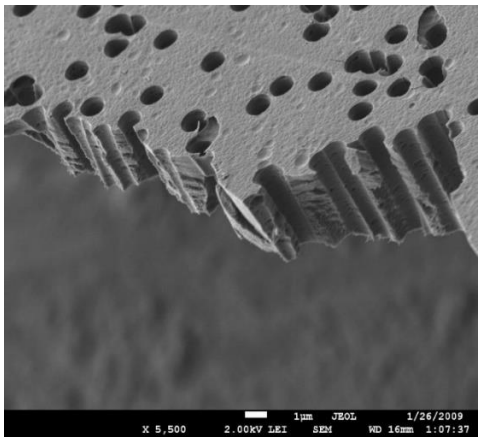


Fig. 2. *Electron micrograph of a cleaved track membrane made of PETP with through cylindrical pores of 1 μm diameter.*

polymer layer becomes equal to a quarter-wavelength phase plate.

Consider the table values of the optical constants in X-ray spectral region [10], limiting ourselves to the elements with a small atomic number Z. These are Li, Be, B, C, Mg, Al, Si, Sc with atomic numbers from $Z = 3$ to $Z = 21$, frequently used as refractive materials for manufacturing focusing optics, thin films, filters and coatings in soft and hard bands of X-ray spectrum. In a wide wavelength range (0.1-10 nm) thin layers of these elements can be considered as phase objects, except the areas near the edge of absorption. Also, out of photoabsorption jumps, the optical constants of these materials have a power wavelength dependence law, the exponent being constant over the entire aforementioned range. It turns out that the intersection points of the curves $\delta(\lambda)$ and $\beta(\lambda)$ on the wavelength axis (in the vicinity of K absorption jumps) set the boundary values of waves λ_c, locating on a single curve $\lambda_c(Z)$ for the elements with different Z (Fig. 1b). The shorter is X-ray wavelength compared with $\lambda_c(Z)$, the easier it is to examine the object with this atomic number by the methods of phase contrast. For a given sample composition, it holds approximately: $\beta/\delta \sim (\lambda/\lambda_c)^2 << 1$, that is, for small wavelengths compared with critical λ_c, the object obviously has phase properties. Note that the equality of real and imaginary additions to unity in the Eq. (1) holds for any Z in some wavelength region near the edge of the absorption band $\lambda_{K_{ab}}$ [10] (see Fig. 1b), therefore $\lambda/\lambda_{K_{ab}}(Z) << 1$ ratio is a sufficient condition to consider given object as purely phase, and not amplitude screen. For substances containing different elements in their spectral formulas we should consider effective (averaged over ensemble) atomic number Z.

3. TEST OBJECTS IN X-RAY MICROSCOPY

As in the hard X-ray phase contrast microscopy of greatest interest for the researcher are transparent three-dimensional objects with characteristic linear dimensions of the order of 10-0.1 microns [7], high requirements are made to the performance of modern X-ray microscopes. This applies above all to the spatial resolution of the device, which should be at several tens of nanometers level over the entire field of view, with the absence of noticeable chromatic aberration. As "phantoms" at the initial setup and research stages in the imaging X-ray microscopy usually simplest test objects are used, whose geometry and size of elements are known from independent sources or gauge measurements. It can be a thin thread of boron or polypropylene fiber, spherical latex or polystyrene microparticles, perforated substrate − sample holder with micro-holes used in transmission electronic microscopy [14], various kinds of microgrids, fragments of a transmission diffraction grating or Fresnel phase zone plate, chips of various microstructures, etc. These micro-objects are used mainly for demonstration purposes, for qualitative microscope performance assessment. Quantitative measurements, such as determination of the image contrast frequency characteristics, are conducted using more

complex tests. Special metrological structures having elements with different controlled sizes are used, such as so-called Siemens star [15] – high contrast radial microstructure, designed to work in a given X-ray spectral band.

In this paper we propose to use as phase test objects for imaging X-ray microscopy porous track membranes (Fig. 2) [1] or their inorganic thin-film replicas, more persistent objects when working with intense sources of X-ray radiation. Track membranes, as structurally heterogeneous samples, satisfy the above requirements presented to the test objects when working in hard spectral region in phase contrast mode. Here we'll consider samples with round cylindrical channels and axes perpendicular to the plane membrane surface, although there are other types of track membranes (with different axis orientation and more complex pore shape [17]). Pore diameter D, varying from 10-20 nm to several microns, depending on the manufacturing conditions, sample can be considered almost constant across the membrane area ($\Delta D/D << 1$), pore walls, to a first approximation, are assumed perfectly smooth. Track membranes with such pore geometry can be considered as reference perforated samples with calibrated through holes randomly distributed over the film area. Pore density N depends on the dose being loaded during the exposing the polymer film to heavy ions and can vary with the top limit of the order of $10^9 - 10^{10}$ sm^{-2} [1]. It is possible to make films with low etched hole density, even with a single micro- or nanohole in a sample with the area of the order of 1 cm^2 [18].

It should be noted that, thanks its unique properties, simple and controlled micron pore structure and submicron track membranes attracted the attention of researchers in various related fields of science and technology from the very beginning of this technology development (see reviews [19, 20]). Besides, the opportunity to localize radiation field or particle flux in areas with very small transversal dimensions – of the order of the track membrane pore cross section, has been used in experimental research (coordinate detector resolution, recording element development in contact lithography [21] and photography [22], spatial resolution control methods, astigmatism compensation in raster translucent microscopes [23], etc.). In X-ray optics, several works appeared in which track membranes were used as strong supporting structures for thin X-ray filters or as selective spectral filters having high transparency in ultra-soft X-ray spectrum but effectively blocking longer wavelength ultraviolet, visible and IR background radiation from the observed object (Sun, laboratory plasma source, etc.) [24-27].

Earlier, an attempt was undertaken to carry out X-ray microscopic studies of track membranes with through pores with moderate (low compared with diffraction limit) resolution in the amplitude contrast mode [28]. And finally, in a recently published work V.I. Balykin with colleagues [29] considered the possibility of subwavelength light localization by passing an excited atom through a nanohole in the track membrane. The issue of fixing the phase of X-ray radiation using track membranes, interesting for applications in phase contrast microscopy, to our knowledge, so far was not studied.

When using the simplest test objects such as fine threads, fine meshes or perforated screens with micro-holes, microscope resolution usually is evaluated "by eye" by contrast images of structural elements of the test object or, more precisely, by blurring its edges. In a track membrane with identical cylindrical through pores that are used to test a microscope, as metrological dimensions of the elements can be considered pore diameter D, average distance between pores $\bar{D} = 0.5N^{-1/2}$, determined by the

pore density *N*, and a variable size of jumpers that appear in the object in case of closely spaced pores or at their intersection. Note that calibrated pore density in a track membrane gives an easy way to estimate magnification of the microscope.

In this paper, we restrict ourselves with consideration of the phase-amplitude evolution of an X-ray plane wave, in submicron spectrum band, passing through a flat layer of a homogeneous material (polyethylene terephthalate) with a single cylindrical through pore. In our simulation, the pore axis was perpendicular to the sample surface, which was considered to be parallel to the incident wave front. The wavelength selected for the calculations corresponds to the characteristic X-ray emission line of copper CuK_{α} equal to 0.154 nm (8.047 keV). Since the spectral properties of all materials in this classic X-ray spectral regions are well known [10], the results obtained using this data are easy to recalculate for other wavelengths and materials.

4. WAVE FIELD CALCULATION IN THE FILM WITH A SINGLE PORE

For numerical simulation of X-ray propagation through a single hole in the membrane we use the method of parabolic equation ("parabolic wave equation, PWE" in English literature). Physical meaning of the parabolic equation approximation has been set out in the pioneering work of Leontovich, Fock and Malyuzhinets [30-33]. This powerful computational tool is widely used in diffraction theory for solving problems of underwater acoustics, radio wave propagation, remote sensing, in fiber optics, etc. [34,35]. Basics of the PWE method and examples of its use are included in textbooks and monographs on electrodynamics and physical optics [36-39]. Taking into account geometry of the object and properties of X-ray radiation, our problem can be considered in the scalar approximation [40-42], so we write down the wave equation for the electric field strength in the following form:

$$\frac{\partial^2 E}{\partial z^2} + \Delta_{\perp} E + k^2 n^2(\vec{r}) E = 0, \quad \Delta_{\perp} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \quad (8)$$

Without repeating classical derivation, recall the basic assessments. In the case of a homogeneous space: $n(\vec{r})$ = const, equation (8) has a solution in the form of a plane wave: $E(x,z) = Ae^{ikn(x\sin\theta + z\cos\theta)}$, propagating at an angle θ to the z axis, x being lateral coordinate.

At small angles θ << 1 (paraxial approximation) the solution can be written as $E(x,z) = e^{iknz} u(x,z)$, where the wave amplitude $u(x,z) \approx Ae^{ikn\left(x\theta - z\frac{\theta^2}{2}\right)}$ has a characteristic oscillation period $\Lambda_{\perp} \sim \frac{2\pi}{kn\theta}$ in the transverse direction x and $\Lambda_{\parallel} \sim \frac{4\pi}{kn\theta^2}$ along the longitudinal axis z. Besides, $\Lambda_{\parallel} \gg \Lambda_{\perp} \gg \lambda$, i.e. $u(x,z)$ is a function slowly varying in the propagation direction along the z axis, namely: $\partial u / \partial z \ll \partial u / \partial x$.

By using this ratio and turning to the case of three spatial variables (*x, y,* z), we obtain from (8) for the wave amplitude an equation of evolutionary type (Leontovich "parabolic" equation [30-32], or "transversal diffusion equation " in Malyuzhints' terminology [33]):

$$2ikn\frac{\partial u}{\partial z} + \Delta_{\perp} E + k^2 n^2 u = 0. \quad (9)$$

In the case of axial symmetry that we are interested in, it is convenient to rewrite write it in cylindrical coordinates:

$$2ikn\frac{\partial u}{\partial z} + \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + k^2 n^2(r,z) u = 0. \quad (9a)$$

Parabolic equations (9-9a) describe propagation of the package electromagnetic waves with wave vectors directed mainly along the axis z. Applicability of the parabolic equation method in a particular field of physics depends on specific conditions: material (optical constant) and the nonuniformity scale of the environment, object geometry and boundary conditions. In X-ray optics, application of

the parabolic equation method for practical problems has been developed in [40-44].

In short-wavelength diffraction and propagation problems, phase of the solution to the wave equation (8): $\arg E(r,z) = kz + \arg u(r,z)$ is a fast oscillation function of longitudinal variable z, while slowly changing argument of the complex wave amplitude $\arg u(r,z)$ characterizes its deviation from the phase of a plane wave e^{iknz}.

Computational advantages of the parabolic equation method, noted in classical works of Malyuzhinets and Tappert [33, 34], are due to transition to slowly varying wave amplitude $u(r,z)$. Instead of a full boundary value problem for the elliptic equation (8), we will solve a Cauchy problem with initial conditions at the entrance to the computation domain. Eliminating the main oscillating factor e^{iknz} sharply reduces computational costs in the short-wave propagation problems. A useful feature of the method is the ability of equivalent truncation of the computational domain by setting exact transparency conditions at its boundaries, resulting from accurate matching the numerical solution with the exact analytical solution in free space [44]. Finite difference approximation of parabolic equation (9) can be performed using absolutely stable implicit six-point Crank-Nicolson scheme [40,45,46]. This approach provides high accuracy and avoids computational error accumulation with distance.

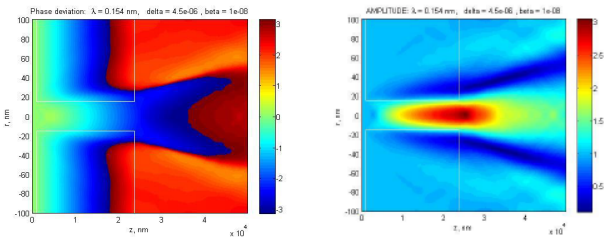
An elementary estimate of the approximation error, following from the comparison of particular solutions of the exact wave equation $E(x,z) = Ae^{ikn(x\sin\theta}$ with their "parabolic" approximation $u(x,z) \approx Ae^{ikn\left(x\theta - z\frac{\theta^2}{2}\right)}$, bounds the distance $z \le a^4/\lambda^3$ where the latter gives accurate amplitude and phase values of the approximate solution. Here $\lambda = 2\pi/k$ is the wavelength of X-ray radiation, a – characteristic transverse object size or non-uniformity scale of the environment; in our case $a = D$. This condition, as well as the estimates of the finite-difference scheme steps $h << a$, $\tau << a^2/\lambda$ are not restrictive in the problems of X-ray optics; besides, there are efficient analytical techniques for increasing computational range and angular sector of the parabolic equation [40-42]. Another inaccuracy, due to backscattering neglection and overcome with the method of coupled waves, is not critical in X-ray optics problems due to small variations of relative refraction index.

5. PHASE VARIATIONS NEAR NANOPORE IN A TRACK MEMBRANE

Consider phase distortion of a plane wave passing through an extremely narrow pore in the membrane. Such blurring limits spatial resolution of a screen with nanoholes used as an X-ray phase-contrast test object. Our calculations were performed by numerical solution of the parabolic wave equation [40,41] using tabular values of the optical constants for the membrane material at the copper CuK_α wavelength (0.154 nm). For polyethylene terephthalate at this wavelength $\delta(\lambda) = 4.5E^{-6}$ and $\beta(\lambda) = 1.0E^{-8}$. Corresponding values of absorption length L_e and quarter-wave thickness phase plate L_φ for homogeneous PETP and the selected wavelength are equal to $L_e = \lambda/4\pi\beta = 1.22$ mm and $L_\varphi = \lambda/4\delta = 8.5$ μm.

Here, we present calculated phase and amplitude of the wave field $u(r,z)$ of the X-ray radiation passing through a single pore of diameter $D = 30$ nm in a film with thickness of $L = 22.5$ μm. For convenience, numerical results are presented in color scale (**Fig. 3***ab*) and contour mode (**Fig. 4**).

Two-dimensional phase and amplitude wave field spatial distribution produced by the object under study allow qualitative assessment of a membrane with nanoholes as a test object for



a *b*

Fig. 3. *Spatial distribution of wave field phase (a) and amplitude (b) in the vicinity of a single 30 nm porein 22.5 μm thick PETP film. The scale is compressed along the pore axis by a factor of 103. Boundaries of the sample and the pore channel are shown by thin light line.*

X-ray microscope. More detailed quantitative information on the test object and functional possibilities of nano-perforated films is reached by studying linear plots of phase and amplitude distributions for membranes with different pore diameters and variable thickness. In **Fig. 5-6,** for a selected sample thickness 22.5 μm and a pore diameter of 30 nm, phase distribution and field amplitude along the pore axis and in the transverse direction on the back surface of the sample are presented.

The following **Figs. 7-9** demonstrate how the phase deviations from the plane wave front in the vicinity of a single pore vary when reducing its diameter from the maximum (200 nm), to the average (50 nm) and minimum (10 nm). The selected smallest diameter value approximately corresponds to the smallest pore size that can still be etched

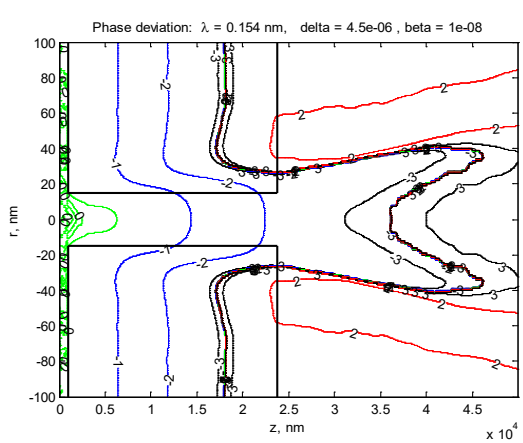
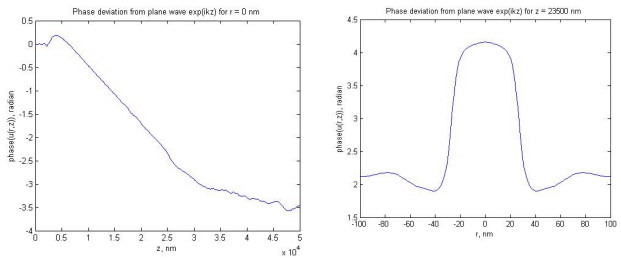


Fig. 4. *Contour representation of the phase shift in a single pore with diameter of 30 nm (compare Fig. 3a).*

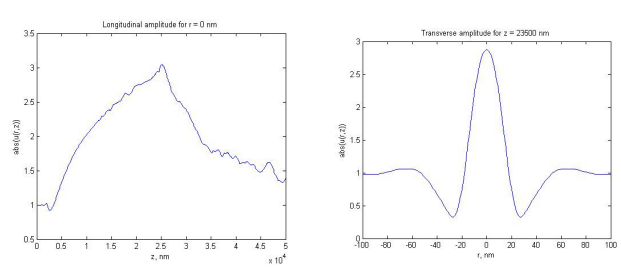


a *b*

Fig. 5. *Axial (a) and radial (b) phase distribution on the perforated membrane back surface. D = 30 nm, L = 22.5 μm. The input end of the sample is shifted from the left boundary (Fig. 3a) by 1000 nm.*

in a thin track membrane by irradiation of the sample with heavy ions. In these figures, spatial phase distribution near the pore presented in different visualization modes: false colors, radial plot on the back side of the sample, grayscale image and contour map of phase distribution − highlight main features of the wave front distortions from the incident wave one.

Numerical calculations of the wave amplitude and phase shift by passing the membrane with a single hole have been made for samples of three thicknesses L = 22.5, 10 and 5 μm and a set of pore diameters *D* in the range from 5 nm to a few micrometers. As expected, for large pores the wave phase portrait at the output face of the membrane reproduces with good accuracy geometry and size of the pore opening, while for small diameters, starting from some threshold, transversal phase blurring of the wave packet was observed near and beyond the channel



a *b*

Fig. 6. *Axial (a) and radial (b) amplitude distribution on the back surface. D = 30 nm, L = 22.5 μm. The input end of the sample is shifted from the left boundary (Fig. 3a) by 1000 nm.*

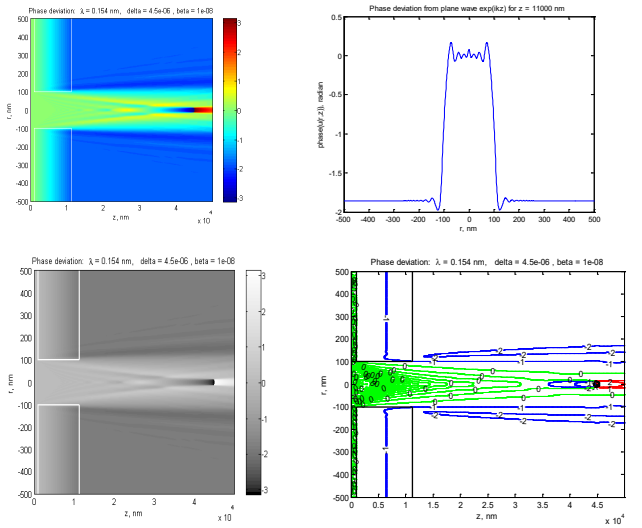


Fig. 7. *Spatial phase distribution of X-ray beam passing a pore of 200 nm diameter in a PETF sample of 10μm thickness.*

outlet. For extremely small pores, the diameter of the phase perturbation zone was more than by an order of magnitude higher than *D*. To find the functional dependency of the phase blur values on the pore parameters pores of the test sample (i.e., *L* and *D* values), turn attention to the plots, where numerical data of the phase blurring on the rear side of membranes with different thicknesses (**Figs. 10-11**). For the sake of easier analysis of the calculation results in different variable ranges, the data are presented in different scales: in relative (normalized to

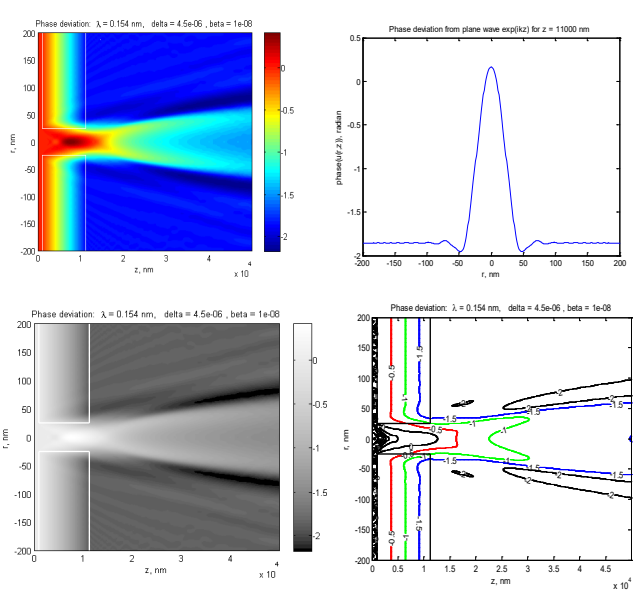


Fig. 8. *Phase distribution near the pore of 50 nm diameter for a sample 10 μm thick.*

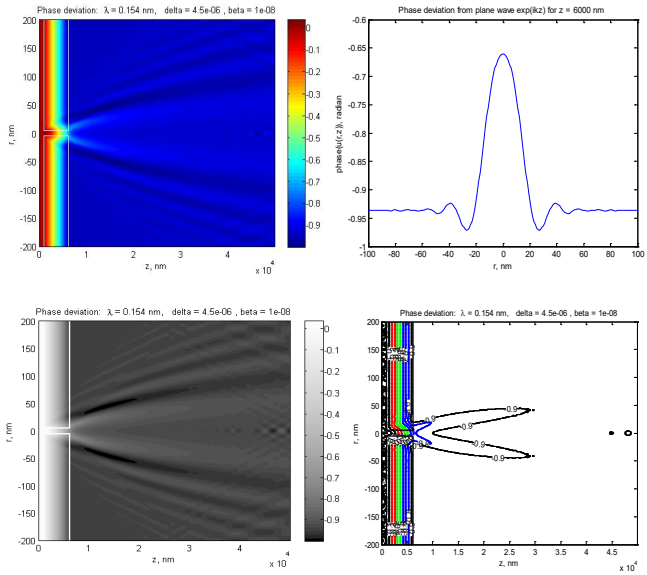
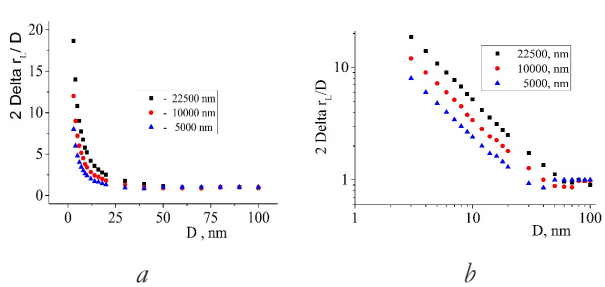


Fig. 9. *Phase distribution near the pore of 10 nm diameter for a sample 5 μm thick.*

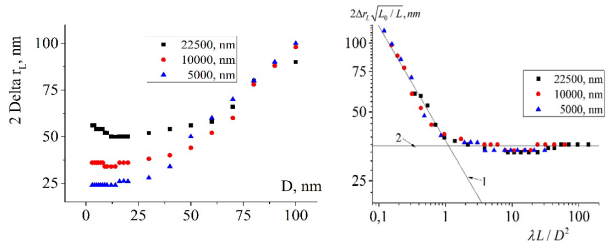
pore diameter, Figs. 10 *a, b*) and absolute form (Fig. 11 *a*).

For each of the presented plots, three regions can be distinguished where the phase blurring near the exit of the channel behaves differently with changing pore diameter. When *D* value is large − about 0.05 μm and more, the phase pattern of the hole in the transverse direction is equal to this diameter, the transmitted wave front is flat, except small edge effects), and the average phase coincides with the free-space phase of the primary plane wave. It is clearly seen in the color Fig. 7a - the exit aperture looks monochrome and its "color" matches the "color" of the incident wave phase. In this case, radiation impinging on the sample, diffracts only at



a *b*

Fig. 10. *Diameter of the phase jump region at the end of the sample, measured at half maximum of the axial phase plot (Figs. 7-9, b) in linear (a) and logarithmic scales (b), for three different samples of a thickness of 22.5, 10 and 5 μm.*



a *b*

Fig. 11. *Absolute width of the phase blur region, as a function of pore diameter, for samples of different thicknesses (a); same, normalized to the square root of the film thickness, depending on the wave parameter (inverse Fresnel number) (b). By condition, =10000 nm (b).*

the pore inlet edges and almost all radiation "falls" into the channel, weakly interacting with the pore walls. Another behavior of phase curves, due to propagation in the sample material is observed for small pore diameters ranging from almost zero values to several tens of nanometers. In this case, the size of the phase blur region at the output plane of a given sample thickness is almost constant, independent of the hole diameter *D*. With changing thickness of the membrane *L*, the transverse phase blur diameter increases approximately as a square root of *L*. A transition region between these two modes is observed, when so-called diffraction beam length D^2/λ is approximately equal to the channel length *L*, i.e., the dimensionless wave parameter $\lambda L/D^2$, the inverse Fresnel number, of the order of unity. In our problem, in normalized variables $X = \lambda L/D^2$ and $Y = 2\delta_L\sqrt{L_0/L}$, all numerical values of the phase blur on the back side of samples with different thicknesses are described by a single curve (Fig. 11*b*). In this figure, lines 1 and 2 are drawn − linear interpolation results for large (1) and small (2) pores, outside the transition area. Curve 1 corresponds to the law $2\Delta r_L = D$, valid for pores with diameters about 0.1 μm and more, while the horizontal line 2 describes the fact that for small pore diameters phase localization (or blurring) at the back of the sample near the nanohole is independent of **D** and grows as *L* with the increasing film thickness. Note that for all our samples, the angular size of the region of localization of the phase perturbation is equal by the order of magnitude to $2\Delta r_L / L \approx \theta_{cr} = \sqrt{2\delta} = 0.003$, where $\theta_{cr} = \sqrt{2\delta}$ is the critical angle of the external total reflection from the nanohole walls for a given wavelength.

Analyzing the results of calculations presented in Fig. 11b and returning to the original problem statement, it can be stated that a track membrane with nanoholes, as a test object for phase-contrast X-ray microscope, satisfies, with a margin, modern spatial requirements for these devices [7], even in near-field mode, when working with relatively large test samples pores, i.e. in the case of large Fresnel numbers. The plots of phase blurring in Figs. 10-11 allow one to choose for the experiment an appropriate track membrane for testing X-ray microscope with necessary spatial resolution.

It should be noted as well that in this problem the main variable determining the solution is Fresnel number of the pore, of course, taking into account frequency dispersion of the sample material. The role of tilting the membrane with a nano-hole from the beam direction is not so important as we consider relatively thin test objects. It is similar to the interference colors of fine soapy or polymer films when the color sequence is weakly dependent on the orientation of the film relative to the light source and observer.

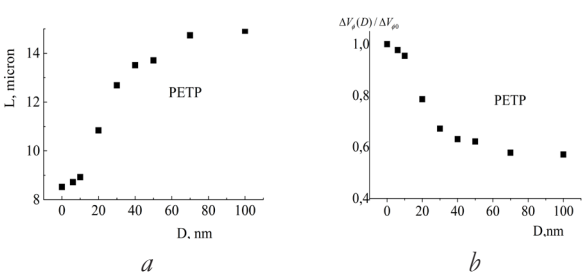
From Fig. 11b and requirement not to be too low-contrast object we can infer that an optimal by spatial resolution phase screen must have through pores whose dimensions satisfy the conditions $\lambda / D \sim \theta_{cr} = \sqrt{2\delta}$ or $D / L \sim \theta_{cr} = \sqrt{2\delta}$, then the same order of magnitude, as noted above, there will have

angular wavefront blur equal to $2\angle r_L/L$ at the pore exit.

6. PHASE VELOCITY OF WAVES PROPAGATING ALONG THE AXIS OF CYLINDRICAL NANOHOLES

In large pores of a track membrane X-ray radiation propagates with light velocity, like in free space. In homogeneous membrane material, phase velocity is greater: $v_{\phi 0} = c/1-\delta \approx (1+\delta)c$. With increasing pore diameter D from several nanometers to the values of about 0.1 μm, phase velocity $v_\phi(D)$ increases from $v_{\phi 0}(D)$ to the light velocity in vacuum c. Using parabolic wave equations, we can easily determine this relationship: $v_\phi = v_\phi(D)$ for a selected wavelength. From color plots of the phase deviation near the nanopore, by varying diameter D, we find the length $l_{\pi/2}$ where the phase shift versus plane wave in free space equals $\pi/2$ (**Fig. 12***a*). In other words, we define a quarter-wave thickness on the axis of the perforated plate and determine its dependence on the pore diameter. Based on these results, one can calculate relative wave deceleration along the channel axis with increasing its diameter D: $\Delta V_\phi(D)/\Delta V_{\phi 0} = V_\phi(D) - C/(V_{\phi 0} - C)$ (see Fig. 12*b*).

From this model problem of X-ray transmission through a material layer with a narrow cylindrical channel, we can conclude that porous track membranes with parallel pores is capable to act as a transparent phase screen providing spatial phase modulation. Fine phase modulation in the sample plane allows one to use track membranes as inhomogeneous phase test objects with deep phase modulation with spatial frequencies up to $2\cdot10^{-2}$-10^{-3} nm^{-1}, which corresponds to the spatial resolution requirements of modern X-ray phase contrast microscopy [7].



a *b*

Fig. 12. *Thickness of the effective quarter-wave layer along the axis of the pore depending on its diameter (a) and corresponding reduction of phase velocity in a hollow duct with its growing width D(b).*

7. TRACK MEMBRANE AS X-RAY DIFFUSER – FILTER FOR SPECKLE SUPPRESSION

Slightly absorbing X-ray frosted screen is an optical an analogue of an ordinary thin matte plate widely used in the visible spectrum band ("frosted glass", "ground glass") for visualization of the optical image, changing the light field characteristics in lighting systems, in various kinds of light diffusers or as a way to control coherent properties of the incident laser beam. In the X-ray band, there is also demand for the use of matte reflective surfaces, frosted screens and transparencies, not only as test objects but also to reduce the grazing angle specular beam reflection, to change spatial coherence of synchrotron radiation, for suppressing speckles in imaging optics, etc. [47-49]. The aforementioned numerical results on the light wave transmission through a single nanopore film allow one to draw some conclusions on the phase of the wave passing through an ensemble of randomly distributed pores, i.e. for a realistic track membrane, not a single pore model sample.

A track membrane with relatively large pores, when the Fresnel number of the pore cavities is greater than unity ($D^2/\lambda L \geq 1$), for hard X-rays presents a weakly absorbing transparency modulating the wave phase with a shift $\Delta\phi = 2\pi\delta L/\lambda \simeq 1$ in the pore openings compared to the uniform layer of membrane. The fraction of the sample area where the transmitted wave experiences phase shift

commensurate with the membrane "nominal porosity" $P_N = \pi D^2 N/4$, where N is the pore density. For highly porous membranes, taking into account the mutual pore intersection, effective membrane porosity is equal to $P_{eff}(P_N) = 1 - e^{-P_N}$ [50]. Therefore, a stack of identical track membranes consisting of $n(P_N) \geq 1/P_N$ layers laid "back to back" serves as an X-ray diffuser, destructing transversal coherence of the incident beam. In order to not to have interfering radiation transmitted through different pores for large transverse correlation length of the incident beam, one can use highly porous track membranes with tilted pores and wide tilt angle distribution that are usually made for filtering gases or liquids. Such membranes usually have the tilt angles of the pore axes (two- or one-dimensional) within ±30 degrees to the membrane surface. Tilted pores provide reduced role of mutual pore crossings and improved membrane performance in standard filtering problems [1]. With a reasonable number of layers about 5 to 10, a stack of porous membranes with a thickness of 5-10 μm still weakly absorbs the radiation and can be considered as a perfect matte screen for hard X-rays -n the through mode. This type of structural heterogeneous filter is now in demand in experimental X-ray microscopy with powerful 3rd or 4th generation X-ray sources.

According to estimates, track membranes with extremely narrow pores and small Fresnel numbers can play the role of ideal matte phase screens transparent for high spatial frequencies even in single- or double layer options if the pore density and diameter satisfy the requirements $N \geq 10^{10}$ sm^{-2}, $D \leq 50$ nm. Note that high-density small-pore samples with wide angular distribution of the axes relative the sample surface are effective and most preferred for the use as a device called X-ray speckle suppressor [51].

8. CONVERSION OF PHASE CONTRAST TO AMPLITUDE

Another interesting feature observed in our model problem of penetrating of a flat X-ray wave through a film with a narrow single pore is the conversion of phase contrast in the amplitude variations in the sample itself and its surroundings. Figs. 3*b* and 6*a,b* illustrates this effect. An essential amplification of the field amplitude inside and behind the nanohole near its axis takes place for well-known diffraction effects. First, due to the difference in refractive indices refraction occurs, and a wide channel can capture in the pore a significant part of the input radiation flux. Secondly, due to interference of the wave propagating inside the pore, with a plane wave passing through the material sample, not only phase disturbance occurs but also spatial redistribution of the wave field amplitude. Because of the presence of the nanohole, the plane wave impinging onto the film becomes inhomogeneous in the sample and behind it − phase and amplitude spatial modulation occurs. Quantitatively, the amplitude modulation effect by a nanohole can be described by calculating the transmission coefficient of radiation through the pore, defined as the integral of the squared field amplitude over the pore cross-section, with appropriate normalization by the incident power stream. Note that the wave field has the form $E(r,z) = e^{ikn_z}u(r,z)$, therefore this definition is correct for transmitting holes although, generally speaking, a high value of the field amplitude does not always mean an intense radiation flux, as it is, for example, in the case of evanescent waves at the edge of a hole or wave field inside a volume (closed) resonator.

Let us turn to **Fig. 13** illustrating numerical results that show the appearance of amplitude

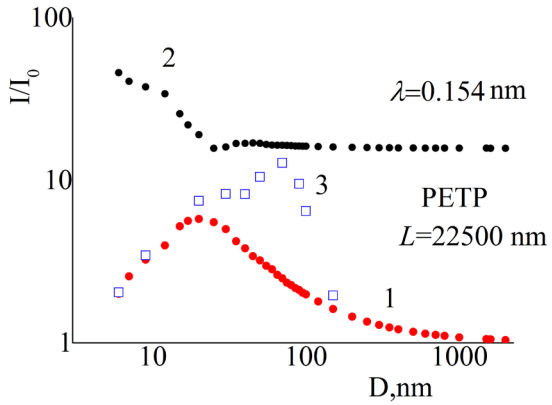


Fig. 13. *Wave field concentration in the pores, depending on their diameter. Curve 1 – normalized energy flow I_0 at the pore entrance, I – energy flow through the exit pore end. Curve 2 – normalized flow integral over the circle of diameter 4D at the output end of the sample. Squares (3) show the dependence of the peak radiation intensity on the pore axis, calculated via maximum field amplitude and normalized by the incident wave intensity.*

contrast on the sample and the increase of the transmission coefficient in a lavsan film with a diameter in the nanometer range. The points of Curve 1 show how the pore transmittance, determined in a standard way (power ratio past stream to that falling on the input aperture) depends on the pore diameter. For very small (several nanometers in diameter) and large (micron-sized) pores, transmittance does not differ much from one, as could be expected. But for the samples with intermediate values of diameter, the transmittance curve I/I_0 has a pronounced maximum at $D \approx$ 20-30 nm. In this case, the radiation flux density through the pore, averaged over its section, at the output end of the membrane by an order of magnitude exceeds the flow density in the incident cross section. For large pores with diameters of the order of 1 μm or more, phase conversion in the amplitude contrast on track membrane models is not observed.

The points of Curve 2 in Fig. 13 are plotted by summation radiation flux at the membrane output around the nano-hole over a circle of diameter $4D$ with the same normalization procedure. Because the membrane material itself with such thickness L absorbs almost nothing, the large part of Curve 2 corresponds to a constant value $I/I_0 = 16$. But for small values D, an increase of transmission curve with decreasing diameter is observed that shows what area of the entrance surface around the nanohole begins to capture the incident wave due to refraction and diffraction at the entrance edge of the pore. Line 3 in the figure, depicted by squares, shows the maximum local values of the field amplitude squared modulus on the pore axis, depending on its diameter and normalized to the input value.

The appearance of amplitude contrast with uniform illumination of the sample in this task, due to phase modulation of the wave front near the nanohole, accompanied by the radiation flux density spatial redistribution near the pores does not contradict with the law of energy conservation.

This article does not cover wave field evolution (phase and amplitude) when propagating further behind the sample. But on the basis of the above results it can be stated that the appearance of amplitude contrast assures visualization of the phase object with nanopores as a test transparency for X-ray microscope calibration.

It can also be noted that a high X-ray flux density provided by nanopores with allows, in principle, to achieve a simple way of contact X-ray lithography (without preparing special templates) and make replicas of track membranes with deep and extremely narrow pores (tens of nanometers in diameter).

9. INORGANIC TRACK MEMBRANES AND THEIR REPLICAS AS PHASE TEST OBJECTS

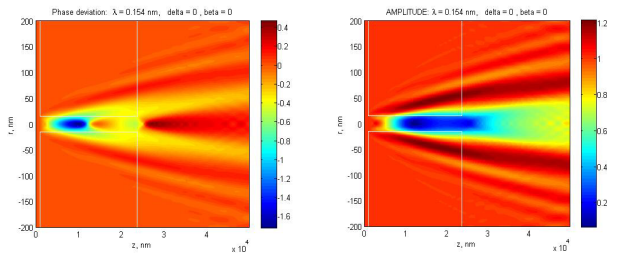
One of the weak points of phase objects made of polymer film is their relatively low thermal and radiation durability. Organic polymeric

materials, even those having benzene rings (cyclic polymers) and crosslinked macromolecules with bulk bonds with the formation of a net structure have too low heat and radiation resistance to serve as material for power optics and function in intensive beams of X-ray (especially, synchrotron) radiation. However, with moderate fluxes of ionizing radiation PETP films have long been used in X-ray experiments, mainly as filter material, substrates or windows of radiation detectors. Simple estimates show which limiting intensity of the incident beam and what absorbed radiation dose allow working with a test object made of a PETP polymer film. With monochromatic illumination at a wavelength of 0.154 nm, the admissible X-ray flux density by uniform illumination of a 22.5 μm thick test sample lies within 10 to 20 W/sm^2. This assessment takes into account that a film of this thickness only absorbs about 2% of the incident radiation, and heat loss of the sample occurs with large efficiency in the near infrared spectral band from both sides of the film [52]. Due to radiolysis of PETP under an ionizing beam with such a limiting for this polymer flux density [53], the "radiation" of the considered test object is about 1 minute (excluding temperature effects). Really, in the image registration schemes of X-ray microscopes with CCD coordinate receivers, the flux density is less than this limit values by orders of magnitude. Therefore, prolonged operation of the proposed in this article polymer test object is possible without significant deterioration of its performance. Anyway, such a test object from PETP film, due to more efficient heat transfer, is not so vulnerable under X-ray beam as most biological microstructures under study [54].

In conclusion, we make the following brief comment. Thin film porous objects with topology typical for track membranes are manufactured by track technology (including nano-porous objects) not only from organic polymers, but also basing on inorganic substrates (glasses, mica, oxide films, etc. [55]), more suitable for experiments with intense X-ray beams, than polymer objects. It can be also deep replicas, including thermo- and radiation resistant metal or ceramic structures obtained by using polymer track membranes as source matrices (templates) with narrow cylindrical or conical pores [56,57]. Strictly speaking, the term "replica" of the track membrane is now applied in a broader sense than was stated in technical dictionaries: "An exact copy of the sample surface on which the structure of the material is clearly expressed".

Track replica manufacturing method appeared almost simultaneously with the birth of track technology [54]. Metallic or carbon thin-film replicas of the surface of the samples with the tracks of fast heavy ions in a solid massive sample or polymer film previously were intended mainly for density measuring density of the tracks and their lateral dimensions with the use of electron microscope. From some time, the researchers learned to make volumetric replicas of the tracks etched not only on the surface, but throughout the depth of the sample irradiated with ions [55,56]. At present, various secondary structures based on deep replicas of track membranes ("template method") received wide spread in various technical and applications as substantive functional elements in nanotechnology [1,57].

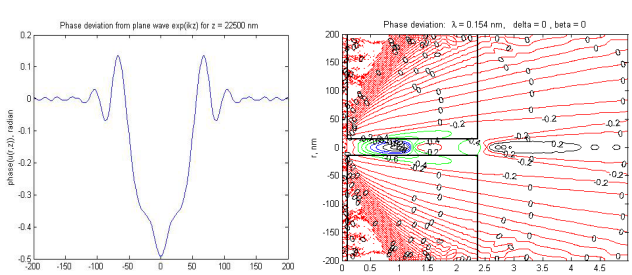
Replicas of track membranes, as phase objects of X-ray optics, can be of three types. The first type is a thin-film surface replica (no bulk structures), providing phase variations much less than unity (low contrast phase test objects capable to model some properties of biological micro-objects by X-ray exposure).



a *b*

Fig. 14. *Spatial phase distribution (a) and wave field amplitude (b) in the vicinity of a cylindrical nano-rod made of PETP, with diameter of 30 nm and length of 22.5 μm.*

Of the second kind are deep (volumetric) replicas reproducing to some degree the structure of the original track membrane template (positive replica) [58,59]. And finally, the third kind reversed (negative) copies in which hollow pores of the track membrane are replaced by a substance for example, metal [55,56] or dielectric (cyacrine glue allowing to realize one of the simplest ways to make bulk replicas from track membranes [60, 61]). Strictly speaking, negative replicas are geometrically more similar to biological microstructures than track hollow membranes themselves with micro- and nanopores. For replicas of this type, X-ray propagation through a nano-column or nano-tip essentially differs from the radiation transport through a hollow pore of the same size, due to the differences in the refractive indices of the material structures. This fact is illustrated in **Figs. 14 and 15** versus the results presented in Figs. 3-6. We can compare the



a *b*

Fig. 15. *Plot of the phase distribution in the rear plane of the sample (a) and the phase contour map (b) around a single cylindrical nano-rod of PETP with diameter of 30 nm and length of 22.5 μm.*

amplitude and phase of the diffracted light at a wavelength of 0.154 nm for two objects of the same shape. In the former case (Figs. 3-6) it was a single hollow cylindrical pore with diameter of 30 nm in a thick PETP film. The latter case presents a similar but reverse (negative) structure: PETP nano-rod with the diameter of 30 nm, length of 22.5 μm and the same orientation relative to the incident wave (Figs. 14-15). The hollow pore serves as an X-ray concentrator amplifying by several times wave field amplitude at the output rod end and providing a noticeable phase shift, of the order of 0.4 radian. In the case of polymer nano-rod (Figs. 3-6), radiation is not being concentrated, but noticeably scattered into free space out of the rod, and at the exit end of the element we see a considerable (almost five times) decrease in amplitude. In the spatial phase pattern, only a weak and localized in the nano-rod cross section phase modulation is seen (by fractions of radian). For the same reason, a biological or other micro-object wrapped in a thin film looks more contrast in phase-contrast X-ray microscopy, provided the real part of the film refractive index is greater than the index of micro-object under study.

Therefore, volumetric positive replicas of track membranes, as test objects for X-ray microscopy, have significant advantages over with similar negative samples. Besides, secondary replication of a negative replica allows one to produce another positive copy – analogue of the primary track membrane, but with a matrix made of a better material suitable for operations with intensive X-ray beams.

10. CONCLUSION

The problem of propagation of X-ray radiation through a weakly absorbing film with a cylindrical through nano-pore with the length of a few microns. This model object is considered as a fragment of a polymer track membrane proposed to be used as a phase

contrast test sample for an X-ray microscope working in a hard region of the spectrum. By numerical solution of the parabolic wave equation, phase shift and amplitude distribution of X-ray radiation penetrating polymer material in the sample and free space near a nano-pore. It is shown that transverse phase blurring at the output end of the membrane can be approximately described by a universal function of the dimensionless Fresnel number defined for a single pore. The issue of phase-to-amplitude conversion in the membrane material and pore cross sections at is studied for various hole diameters. The obtained results on the phase shift magnitudes and their spatial localization near the pore channels allow us to propose track membranes not only as test objects but also as X-ray diffuser or speckle suppressor in imaging optical systems with coherent or partially coherent light sources illuminating micro-objects under study.

REFERENCES

1. Apel PYu, Dmitriev SN. Track membranes. In: *Membranes and Membranes technologies*. Yaroslavtsev AB (ed.), pp. 117-160. Moscow, Nauchny mir Publ., 2013.
2. Zernike F. How I discovered phase contrast. *Science,* 1955, 121:345-349.
3. Zernike F. *Phys. Zeits.*, 1935, 36:848; *Zeits. f. Techn. Phys.*, 1935, 16:454.
4. Shpolsky EV. *UFN,* 1947, 32:376 (in Russ.).
5. Rytov SM. *UFN*, 1950, 41:425 (in Russ.).
6. Franson M. *Phase-contrast and interference microscopes.* (Russian transl. ed. GG. Slyusarev) Moscow, GIFML Publ., 1960, p. 180.
7. Lider BB. *UFN*, 2017, 187(1):201 (in Russ.).
8. Blokhin MA. *X-ray physics*. Moscow, GITTL Publ., 1957, p. 518.
9. James R. *Optical Principles of X-ray Diffraction* (Russian transl. ed. VI. Iveronova). Moscow, Inostr. Lit. Publ., 1950, p. 572.
10. X-Ray Interactions with Matter, 2010. URL: www.cxro.lbl.gov/optical_constants/.
11. Migdal AB, Krainov VP. *Approximate Methods of Quantum Mechanics*. Moscow, Nauka Publ., 1966.
12. Spiller E. X-Ray Optics. In: *Encyclopedia of Optical Engineering,* Taylor & Francis, 2003.
13. Artioukov IA, Kasyanov YuS, Kopylets IA et.al. *Rev. Sci. Instrum.*, 2003, 74:4964.
14. Falch KV, Lyubomirsky M, Casaru D et al. *Ultramicroscopy*, 2018, 84:267.
15. Rehdeun S, Guttmann P, Werner S, Schneider G. *Optics Express*, 2012, 20(6):5830.
16. Mitrofanov AV, Apel PJ, Burmistrov AA, Levkovich NB, Orelovich OL. Interchangeable windows and filters of X-ray image detectors based on porous films of submicron thickness. *Proc. XVth Int. Symp. "Nanophysics and Nanoelectronics",* 2, p. 607. Nizhny Novgorod, Lobachevsky Univ. Publ., 2011 (in Russ.).
17. Apel PYu, Blonskaya IV, Didyk AYu, et al. *Nuclear Instrum. and Meth. In Phys. Research B*, 2001, 179:55.
18. Apel PJ, Didyk AJ, Kuznetsov VI. *Pribory i tekhnika experimenta*, 1988, 6:48 (in Russ.).
19. Flerov GN. *Bulletin of the Academy of Sciences of the USSR*, 1984, 4:35 (in Russ.).
20. Spohr R. *Ion Tracks and Microtechnology. Principles and Applications*. Braunschweig, Vieweg, 1990, p. 272.
21. Valiev KA, Velikov LV, Dushenkov SD, et al. *Microelectronics*, 1982, 11(5):447 (in Russ.).
22. Mitrofanov AV, Shpolsky MR. Porous membrane as a test object for assessment of photo material structural properties. In: *"Prospects of development of photographic registration tools information for astronomical observations. Abstracts"*, p. 20. Dushanbe. Donish Publ., 1983. (in Russ.).
23. Mitrofanov AB, Filippov MN. *Izvestiya RAN, ser. Fiz.*, 1992, 56 (3):112 (in Russ.).
24. Mitrofanov A, Pudonin F, Starodubzev N, Zhitnik I. *Proc. SPIE*, 1998, 3406:35.

25. Mitrofanov AV, Apel PYu. *Nucl. Instr. and Meth.*, 2006, B 245:332.
26. Mitrofanov AV, Apel PYu. *Izvestiya RAN, ser. Fiz.*, 2009, 73(1):61 (in Russ.).
27. Dominique M. et al. *Applied Optics*, 2009, 48(5):834.
28. Andreev AV, et al. *Membranes (ser. Crit. Technol. VINITI)*, 2005, 3(27):17 (in Russ.).
29. Afanasieva E, et al. *JETP*, 2017, 152, 3(9):438 (in Russ.).
30. Leontovich MA. *Izvestiya AN USSR, ser. Fiz.*, 1944, 1:16 (in Russ.).
31. Leontovich MA, Fock VA. *J. Phys. USSR*, 1946, 10:13.
32. Fock VA. *Problems of diffraction and propagation of electromagnetic waves*. Moscow, Sovetskoe Radio Publ., 1970, 476 p. (in Russ.).
33. Malyuzhinets GD. *UFN*, 1959, 69(10):320 (in Russ.).
34. Tappert FD. Parabolic equation method. In: *Wave Propagation and Underwater Acoustics*. Moscow, Mir Publ., 1980.
35. Prokopovich DV, Popov AV, Vinogradov AV. *Kvantovaya elektronika*, 2007, 37(9):873 (in Russ.).
36. Katsenelenbaum BZ. *High-frequency electrodynamics*. Moscow, Nauka Publ., 1966, p. 240.
37. Weinstein LA. *Electromagnetic waves*. Moscow, Radio i Svyaz Publ., 1988, p. 440.
38. Levy M. *Parabolic Equation methods for Electromagnetic Wave Propagation*. London, IET, 2000, p. 336.
39. Rautian SG. *Introduction to physical optics*. Moscow, Librocom Publ., 2009, p. 256.
40. Kopylov YuV, Popov AV, Vinogradov AV. *Optics Communications*, 1995, 118(5-6):619.
41. Popov AV, Vinogradov AV, Kopylov YuV, Kurokhtin AN. *Numerical simulation of X-ray diffractive optics*. Moscow, A&B Publ. House, 1999, p. 29.
42. Attwood D. *Soft X-rays and Extreme Ultraviolet Radiation*. Cambr. Univ. Press, 2000.
43. Bukreeva I, Popov A, Dabagov S, Lagomarsino S. *Phys. Rev. Letters*, 2006, 97:184801.
44. Popov AV. *Radio Science*, 1996, 31(6):1781.
45. Popov AV. *Comp. Math. and Math. Phys.,* 1968, 8(5):1140 (in Russ.).
46. Marchuk GI. *Methods of Computational Mathematics*. Moscow, Nauka Publ., 1977.
47. Matsuura Y, Yoshizaki I, Tanaka M. *J. Appl. Cryst.*, 2004, 37:841.
48. Goikhman AYu., Lyatun II., Ershov PA, et al. Highly porous nanoberillium for suppressing the speckle of X-ray radiation. *Proceedings conf. "X-ray optics-2014"*, p. 29. Chernogolovka, IPTM Publ., 2014 (in Russ.).
49. Naulleau PP, Liddle JA, Salmassi F, Anderson EH, Gullikson EM. *Appl. Optics*, 2004, 43(28):5323.
50. Riedel C, Spohr R. *Rad. Eff.*, 1979, 42:69.
51. Goikhman A, Lyatun I, Ershov P, Snigireva I, Wojda P, Gorlevsky V, Snigirev A. *Journal of synchrotron radiation*, 2015, 22(3):796.
52. Mitrofanov AV. *Kvantovaya elektronika,* 2018, 48(2):105 (in Russ.).
53. Zhitariuk NI, Fiderkiewicz A, Buczkowski M, et al. *European Polymer Journal*, 1996, 32:391.
54. Solem JC, Baldwin GC. *Science*, 1982, 218:229.
55. Fleisher RL, Price PV, Walker RM. *Tracks of charged particles in solids. Part 1. Methods of research tracks*. Transl. from English edited by Shukolyukova YuA. Moscow, Energy Publishing House, 1981, p. 40.
56. Spohr R. Method for producing planar surfaces having very fine peaks in the Micron range. *United States Patent № 4*, Jul.6, 1982, 338, 164.
57. Vetter J, Spohr R. *Nucl. Instrum. Methods in Phys. Res.,* 1993, B 79:691.
58. Toimil-Molares ME, Beilstein J. *Nanotechnol.*, 2012, 3:860.

59. Mitrofanov AV. *Isvestiya AN USSR, Ser. Fiz.*, 1993, 57(8):176 (in Russ.).
60. Mitrofanov AV, Tokarchuk DN, Gromova TI, Apel PYu, Didyk AYu. *Radiation Measurements,* 1995, 25(1-4):733.
61. Apel PYu, et al. *Phys. Chem. Chem. Phys.,* 2016, 18:25421.

DOI: 10.17725/rensit.2020.12.191

Photocurrent Domain Instability In High-Resistance Tunnel CdZnTe-Based Mis Structures

Yuri N. Perepelitsyn

Kotelnikov Institute of Radioengineering and Electronics of RAS, Saratov Branch, http://www.cplire.ru/rus/sfire Saratov 410019, Russian Federation

E-mail: olga-optics@yandex.ru

Received November 21, 2019; peer reviewed December 15, 2019; accepted December 20, 2019

Abstract. **The results of experimental studies of the photocurrent domain instability arising under illumination in high-resistance tunnel MTISTIM structures based on CdZnTe single crystals are presented. It is shown that the photocurrent domain instability is based on drift nonlinearity, i.e., the photostimulated spatial rearrangement of the electric field. It was found that, as in the classical Gunn diodes, the appearance of microwave oscillations of the photocurrent occurs at threshold values of external macro parameters, the change of which within certain limits provides a reversible change in the frequency of the oscillations up to 8 octaves. The results of experimental studies of the velocity–field dependence in CdZnTe, measured under spatially inhomogeneous distribution of the electric field in the volume of the MTISTIM structure, are presented. The threshold field of the oscillation occurrence and the maximum velocity of the majority carriers in CdZnTe single crystals are determined. Numerical estimates of the minimum irradiation power and carrier concentration necessary for the appearance of the photocurrent domain in the high-resistance MTISTIM structure of CdZnTe are presented. It is shown that due to the transverse electro-optical Pockels effect, the change in the domain field of the electro-optical characteristics of the semiconductor component of the diode allows the transfer of optical information from the controlling light flux $I_1(x,t)$ to the probe light flux $I_2(x, t)$, transmitted through the structure, i.e., to carry out high-frequency optical modulation of one light flux by another.**

Keywords: **photocurrent domain instability, photocurrent oscillations, threshold field, high-resistance tun-nel MTISTIM structure, photoelectric domain, electric current distribution**.

UDC 621.382.2; 537.222.22

Acknowledgements: The work was performed as part of the State Assignment.

For citation: Yuri N. Perepelitsyn. Photocurrent Domain Instability In High-Resistance Tunnel CdZnTe-Based Mis Structures. *RENSIT*, 2020, 12(2):191-200; DOI: 10.17725/rensit.2020.12.191.

Contents

1. INTRODUCTION

Recently, the studies related to the development of optically controlled active elements, based on which it is possible to create devices that provide the basic types of optical signal processing (modulation, switching, angular deviation, etc.) in the nanosecond and picosecond time scale have become particularly relevant [1-3].

One of the effects that can provide fast and "strong" changes in the electro-optical characteristics of the medium is the Gunn effect, which is based on the intervalley transfer of carriers. Theoretical estimates show that the time of intervalley transfer leading to the formation of strong field domains is $\sim 10^{-14}$ s, which potentially makes it possible to obtain changes in the electro-optical characteristics of the medium with characteristic times of $\sim 10^{-12}$–10^{-14} s [2,4].

The study of the light exposure effect on the parameters of Gunn diodes began almost

immediately with studies of the Gunn effect itself. The experimental results showed that illumination of planar di-odes leads to the control of the oscillation threshold field, the spectrum and intensity of the oscillations, the improvement of coherence, the change in the oscillation frequency, etc. In diodes based on high-resistance compensated single crystals of Ge(Au) GaAs(Cr), ZnTe-CdTe, the current-voltage characteristic (CVC) under illumination becomes N-shaped, and the formation of a section with negative differential photoconductivity (NDPC) on the CVC is accompanied by the appearance of low-frequency oscillations of photocurrent [5,6].

Later on, it was shown [7] that, depending on the backlight intensity, temperature, and other factors, one or another recombination nonlinearity is realized in such diodes, in which the transfer rate of the formed domains is limited by the time of carrier redistribution between the conduction band and the capture levels.

When studying the photoelectric properties of homogeneous high-resistance MTISTIM structures, where M is optically transparent metal electrodes, TI is tunnelthin insulator layers, S is a high-resistance semiconductor, it was found that the change in conductivity that occurs in such structures under illumination is accompanied by spatial changes in the distribution of electrical field $E(x)$ from uniform in the dark to sharply inhomogeneous under illumination. These changes are practically "inertialess" following the change in intensity. As a result, during illumination, the region of the "strong" electric field is spatially localized at the electrode opposite to the illuminated one, and the restoration of the initial dark field E_0 ($E_0 = V_0/L$) created by an external voltage source in the volume of the structure, after the termination of illumination occurs spontaneously during the drain of photoinduced charge to the external circuit [8].

Further studies of the redistribution of electric fields in high-resistance tunnel MTISTIM structures based on compensated and "pure" p-CdTe(Cl) single crystals with a concentration of deep impurity levels $N_t > 10^{15}$ cm^{-3} and shallow impurity levels $N \sim 10^{12}$-10^{13} cm^{-3}, respectively, and n-CdTe(In) single crystals with $N_t \sim 10^{16}$ cm^{-3} have shown that the difference in electric fields formed during the illumination between the non-illuminated and illuminated electrodes in a number of structures can be significant [9]. Accordingly, it was to be expected that under the conditions of high applied voltages V_0 and high illumination intensities, an NDPC region could appear on the CVC of such structures.

The purpose of the work reported here was an experimental study of the excitation conditions for illumination of the Gunn oscillations of the photocurrent in high-resistance tunnel MTISTIM structures of CdZnTe.

2. MATERIALS AND METHODS

We investigated a batch of samples of homogeneous high-resistance tunnel MTISTIM structures based on undoped $Cd_{1-x}Zn_xTe$ ($x = 0.04$) single crystals with a concentration of deep impurity levels $N_t < 10^{13}$ cm^{-3}, resistivity $\varrho > 5 \cdot 10^8$ ohm×cm and the equilibrium bulk concentration of carriers $n_0 \sim 10^6$–10^8 cm^{-3}. The purity of the initial components Cd and ZnTe was no worse than $6N^{++}$. The prepared samples were rectangular parallelepipeds with different distances between the contacts L and the area of the illuminated surface S ~0.1−0.12 cm^2. Metallic Au or PT contacts were chemically deposited on opposite faces of single crystals containing tunnelthin oxide layers. According to ellipsometric measurements, the average thickness of the tunneling oxide layer d_{Ox} was ~10−15 nm.

The samples were illuminated from the side of the negative electrode by pump light pulses with quantum energies $h\nu \geq E_g$, where hν is the quantum energy, E_g is the band gap. The duration, amplitude and frequency of illuminating pulses of various shapes was regulated using an electronic power circuit. Samples were placed between optically transparent electrodes in a stage series-connected in the gap of the center core of the coaxial line. The photocurrent pulses and photocurrent oscillations generated at different powers of the illuminating pulses P_0 and the constant applied voltage V_0 were taken from the load resistance $R_l = 50$ ohm, connected in series with the sample, applied to the input of the oscilloscope and photographed. Based on the obtained images, the parameters of the oscillations were subsequently evaluated. The spatio-temporal characteristics of the electro-optical response generated in the sample under illumination were estimated using the characteristics of optical pulses recorded at the

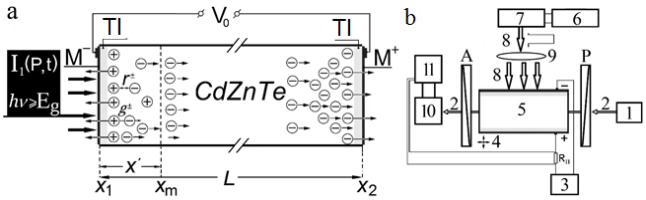


Fig. 1. *a – schematic diagram of the tunnel MTISTIM CdZnTe structure; b – schematic diagram of the experimental setup: 1 - power supply and source of probe light I_2, P - input polarizer, 2 - collimated beam of probe light I2, 3 - high voltage generator; 4 - micropositioner system, 5 - sample, 6 - electric pulse generator, 7 - electronic circuit of the pump light source, 7 - pulses of pump (controlling) light I_1, 9 - micro lens, A - analyzer, 10 - unit for recording optical pulses of probe light , 11 - oscilloscope.*

output of the sample, when it was probed by a narrow beam of non-photoactive light. A diagram of the MTISTIM structure and experimental setup is shown in **Fig. 1**a,b.

3. EXPERIMENTAL RESULTS

The band gap of the studied samples was determined experimentally from the optical absorption curves. According to measurements of several samples at room temperature, the band gap E_g was ~1.52±0.5 eV, and the absorption coefficient was α ~230÷250 cm^{-1}, which agrees well with Ref. [10]. The spectral distribution of photoconductivity obtained using selective spectral filters when the samples are illuminated with the same light fluxes (P ~1 mW) in the spectral range of ~640–1300 nm at a field strength of E_0 ~1.5 kV/cm is shown in **Fig. 2**. It

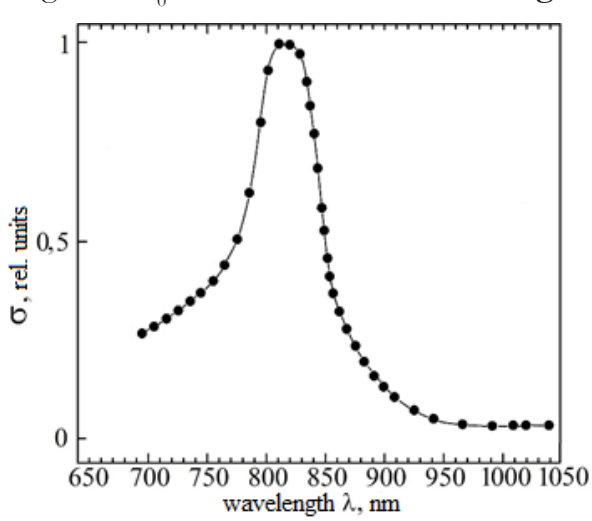


Fig. 2. *Spectral distribution of the photocurrent at the magnitude of the field E_0 = 1.5 kV/cm; T = 300 K.*

is seen that the photocurrent is maximal when the pump light has a wavelength of λ ~0.81÷0.82 μm and decreases sharply when the pump light is shifted to either the short-wavelength or long-wavelength part of the spectrum.

In the initial section, the dark CVCs of the samples are close to linear; up to fields E_0 ~3÷8 kV/cm, a section with a vertical increase in the dark current characterizing the transition from linear to quadratic dependence is not observed. In most samples, in the voltage range V_0 = 0÷800 *V*, the dark current increases by two to three orders of magnitude and does not exceed ~0.6-0.7 μA. At high bias voltages V_0, breakdown of samples occurs at fields E_0 ~13÷15 kV/cm, regardless of the polarity of the applied voltage.

At E_0 > 1÷2 kV/cm and low illumination intensities, the photocurrent increases linearly until saturation. With increasing illumination intensity, a deviation from linearity is observed, and at high illumination intensities, the CVC of the samples has a sublinear form, i.e., in the studied range of applied voltages, there is no carrier injection from the contacts.

Under illumination by light with a power of P_0 ~5–15 mW, a photocurrent appears in the samples, exceeding the dark one by approximately 3–4 orders of magnitude. At a field strength E_0 ~7–9 kV/cm and photocurrents exceeding ~3–10 mA, the CVCs of a number of samples show a spontaneous appearance of an *N*-type NDPC region, the formation of which leads to the formation of δ-shaped or triangular photocurrent oscillations that continue only during the light exposure. In long samples with *L* ~0.15–0.2 cm, with an increase in the illuminating pulse power P_0, the threshold field for the oscillation E_p significantly decreases and increases with a decrease in *L*.

The spectral range of illumination, in which the appearance of oscillations is observed in the samples, is close to the spectral distribution of the photocurrent. However, with the spectral shift of the illumination from the spectral region of intrinsic absorption to the short-wavelength region ($h\nu > E_g$) and especially to the long-wavelength region ($h\nu < E_g$), photocurrent oscillations occur at ever higher values of the external macroscopic parameters, i.e.,

the bias voltage V_0 and illumination power P_0. The occurrence of oscillations ceases when illuminated with light quanta with $\lambda > 0.88–0.9$ μm. In addition, a change in the shape, amplitude, frequency and coherence of the excited oscillations is observed during the spectral shift of the pump illumination.

A study of the kinetics of transients at different relations between the values of external macro parameters and different polarity of the applied voltage showed that:

- the time of photocurrent stabilization depends on the polarity of the applied voltage and is minimal at a positive potential on an non-illuminated electrode;
- in the presence of a natural oxide layer, the regime of the through photocurrent stabilizes at bias volt-ages V_0 ~1.5–3 V which increases with increasing thickness of the tunneling dielectric layer d_{Ox};
- a jump in the photocurrent at the leading edge of the photoresponse, preceding the establishment of a stationary photocurrent, is observed in most samples at low illumination powers ($P_0 \leq 1$ mW) and fields E_0 not exceeding 1.5–2 kV/cm, and the amplitude of the photocurrent jump rapidly decreases with increasing applied voltage;
- at the sub-threshold values of macroparameters, long-term relaxation and spikes of the photocurrent when the illumination is turned on and off are not observed in most samples;
- at any polarity of the applied voltage, an increase in the bias voltage and the power of the illuminating pulse leads to a decrease in the time of flight of carriers and the duration of the leading and trailing edges of the photocurrent pulse, and when the macro parameters are close to the threshold, the shape and time constant of the photo and electro-optical responses approach the duration and shape of the illuminating pulse;
- the NDPC segment appears at a certain power of the optical pulse P_p, below which the generation does not occur at any bias voltage.

Depending on the initial power of the illuminating pulse, two scenarios of initiating oscillation are observed. The first of them is realized at a small excess $P_0 \approx P_p$. In this case, the emergence of the domain is preceded by the formation in the near-contact region of the non-illuminated electrode or on the constant component of the photocurrent pulse of the photocurrent fluctuation, which is preserved while maintaining the macro parameters. Its transformation into a δ-shaped or triangular domain with a domain amplitude close to the amplitude of the photocurrent pulse occurs spontaneously when either of the macro parameters (or both at once) increases by a certain value ΔP or ΔV, which increase with a decrease in L. After the formation of a single domain, further increase in any of the macroparameters with a step ΔP or ΔV (provided that the other macroparameter retains its threshold value) leads to a sequential increase in the number of photocurrent oscillations, generated within the photocurrent pulse. The stabilization of the amplitude and shape of the photocurrent oscillations occurs after the formation of the second domain (**Fig. 3***a*).

The second scenario is realized when $P_0 > P_p$. In this case, the oscillation occurs spontaneously, and the frequency of photocurrent oscillations is determined by the power of the optical pulse (Fig. 3*d*). In this case, regardless of the scenario of oscillation occurrence, the maximum frequency of the oscillations is achieved at certain values $V_0 = V_{max}$ or $P_0 = P_{max}$, which differ from sample to sample. Then the oscillation frequency stabilizes and

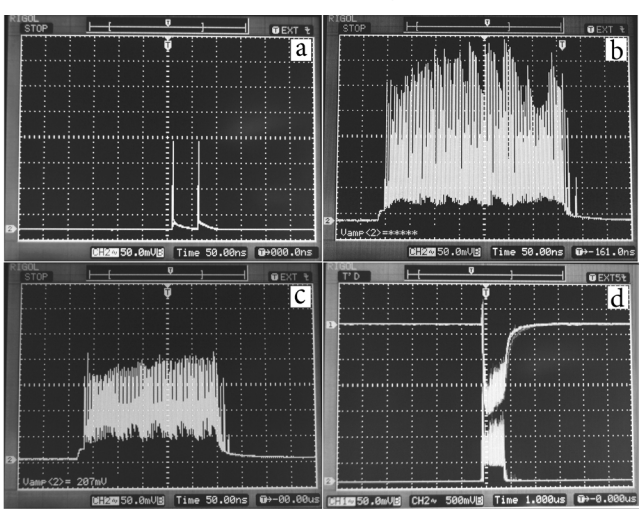


Fig. 3. *Oscillograms of photocurrent pulses generated at various values of external macroparameters: a – $P_0 \geq P_p$, $V_0 = V_p$; b – $P_0 \sim P_{max}$, $V_0 = V_p$; c - $P_0 \sim P_p$, $V_0 = V_{max}$; d - $P_0 >> P_p$, $V_0 \sim V_p$ where the upper trace is the optical pulses of the probe light, the lower trace is the fluctuation of the photocurrent. T = 300 K.*

remains constant, and a further significant increase in any of the macroparameters leads to breakdown of oscillation or to breakdown of the sample.

Accordingly, a decrease in any of the macroparameters in the range from V_{max} or P_{max} to their threshold values is accompanied by a decrease in the frequency of the oscillations and restoration of their shape. When fixing the values of the macroparameters in any interval of their increase or decrease, the oscillation frequency is stably preserved and is determined by the current values of the macroparameters. Changes in the generation frequency with a change in the optical pump pulse power and bias voltage are shown in Fig. 3*b,c*.

In the general case, the frequency of the oscillations is inversely proportional to the distance L between the contacts and weakly depends on the variable macroparameter and the excitation mode, although in some samples the maximum frequency may slightly vary. In most samples with L ~1.5–2 mm, when any of the macroparameters changes in the range of $V_p \div V_{max}$ or $P_p \div P_{max}$, the oscillation frequency changes in the range of 6–8 octaves. With decreasing L, the width of the oscillation band decreases and in samples with L ~100–300 μm does not exceed 2–3 octaves.

The experiments showed that in addition to changing the frequency characteristics, the growth of external macro parameters has a complex effect on the parameters of the oscillations. In particular, one of the manifestations of such an effect is associated with the shape and amplitude of the excited oscillations. Thus, in the majority of the samples studied, the amplitude and period of the formed oscillations successively decrease with the growth of any of the macroparameters, and their shape is successively transformed from δ-shaped to close to triangular or sinusoidal. However, a different situation is realized in a number of samples: with the growth of any macroparameter, up to P_{max} or V_{max}, the δ-shape and amplitude of the oscillations are preserved. In some samples, the formation of oscillations with a nanosecond and shorter leading edge length and domain amplitude close to the amplitude of the photocurrent pulse or exceeding it is observed. In addition, at high optical pump pulse powers, the formation of single "quasistatic" domains with the amplitude not exceeding 0.2 A is observed.

At the same time, although a somewhat lower generation frequency is achieved in some samples with P_{max} or V_{max}, nonetheless, samples in which the initial oscillation parameters are preserved are of significant applied interest, since oscillations of this form provide high-efficiency optical modulation of continuous light fluxes or switching of discrete short optical pulses [11].

Another influence of the growth of external macroparameters is associated with the period and modes of oscillation. Due to this effect, in a number of samples with increasing of any of the macroparameters, an aperiodic transit-time oscillation mode is implemented, in which the period between the oscillations sequentially decreases. With an increase in the macroparameters, a two-domain oscillation mode appears, which transitions into a three- or fourdomain mode as the macroparameter increases, with a close but different frequency of the generated oscillations; there is a spontaneous transition from δ-shaped domains to trapezoidal domains, etc. In addition, under exposure to light pulses of complex shape (sinusoidal, triangular, sawtooth, etc.), in such structures periodic or aperiodic photocurrent oscillation occurs, the envelope of which uniquely repeats the shape of the illuminating pulse [12].

The minimum duration of an optical pulse at which a single domain occurs in the structures exceeds two to three oscillation periods and depends on the power of the illuminating pulse.

4. DISCUSSION OF THE RESULTS

A comparison of the excitation conditions, oscillation conditions, and parameters of the photocur-rent domains with the results of similar studies shows that the formation of photocurrent oscillations in CdZnTe samples is not associated with a decrease in carrier mobility due to optical charge exchange of impurity levels, which is characteristic of recombination instability [7]. The high rate of domain trans-fer, the dependence of E_p on illumination intensity, the aperiodic transit-time oscillation mode, and the reversible dependence of the frequency of oscillations on external macroparameters, observed experimentally, indicate an unambiguous relationship between the formation of N-type NDPCs in

such structures and the photostimulated spatial rearrangement of the electric field, which so far has not been experimentally observed.

The results of experimental and theoretical studies of this effect are presented in Ref. [8], where, based on model concepts, it was shown that for low optical pump intensities and high tunnel transparency of the dielectric layer d_{Ox}, when the diffusion component of the photocurrent and the accumulation of mobile carriers at the non-illuminated electrode can be neglected, the stationary field distribu-tion profile $E(x)$ in the monopolar transport region has the form [8]:

$$E(x) = E_0 \left(\frac{x'}{L} \right)^{1/2}, \quad (1)$$

where $E_0 = \sqrt{\frac{8\pi JL}{\chi\mu}}$, L is the separation between the contacts, $x' = x_m - x_1$ is the width of the region of oscillation, separation and recombination of nonequilibrium photocarriers, x_1 is the initial coordinate, χ is the relative permittivity, μ is the mobility, J is the photocurrent density, determined by the relation [8]:

$$J = \frac{9}{32\pi} \frac{\chi\mu V^2}{L^3}. \quad (2)$$

Dependencies (1) and (2) are expressions known in the literature, discussed earlier in Ref. [13]. They correspond to the so-called "virtual cathode" approximation, from which the "infinitely weak field" draws the carriers off. Correspondingly, dependence (2) is a generalization of the known result for the case of the "photoelectric cathode", when the interband optical pumping is the carrier source. When a homogeneous high-resistance MTISTIM structure is illuminated, the role of such a source is played by a narrow contact layer near the illuminating electrode x', inside which photocarriers are generated, recombined, and separated.

It was shown [14–16] that the magnitude and coordinate dependences $E(x)$ obtained experimentally in MTISTIM structures based on high-resistance "pure" and compensated p-CdTe single crystals do not coincide with the results of theoretical calculations given in [17]. To the greatest extent, these differences relate to the contact areas, where the illuminated electrode has a more significant field decline than theoretical calculations, and the experimental values of the field near the non-illuminated electrode significantly exceed their calculated values and vary greatly with the coordinate [14]. Such differences were associated by the authors of Ref. [14] with the inadequacy of the model of the MSM structure with respect to the real MTISTIM structure, where the through photocurrent is accompanied by a partial accumulation of mobile carriers in the near-electrode regions. It was also shown there that in tunnel MTISTIM structures based on "pure" single crystals, the configuration of the mobile charge accumulated near the non-illuminated electrode determines the coordinate dependence $E(x)$. The latter, along with the dependence $J(I)$, has a sublinear form and can be approximated by the expression $E(x) = A \cdot x^n$, where A is a coefficient, n is the nonlinearity index, $n < 1$ [14].

Experiments on the nature of the distribution of electric fields in CdZnTe structures showed that, at comparable bias voltages and illumination intensities, the difference in electric fields achieved in such structures is greater than in similar structures of Ref. [14]; in the most perfect samples spatial changes of the field in the volume of the structure occur in shorter times. Moreover, at certain relations between the macroparameters, the CVCs of a number of samples show spontaneous formation of a descending N-type region, the occurrence of which is accompanied by the formation of photocurrent oscillations within the photocurrent pulse (Fig. 3*d*).

The steady-state distributions of the field $E(x)$, which are established when one of the macroparameters changes, are shown in **Fig. 4***a,b*. From the coordinate dependences $E(x)$ in Fig. 4*a* measured by the method [16], it follows that the stationary field distributions E(x) in the studied sample and the high-resistance MTISTIM structure based on a "pure" p-CdTe single crystal [14] qualitatively agree, i.e., in both cases under illumination the electric field decreases near the illuminated electrode and increases towards the non-illuminated one, reaching a maximum in its close proximity. However, at comparable bias voltages and power of the optical pump pulses, a stronger field difference between the illuminated and non-illuminated electrodes is formed in the CdZnTe structure.

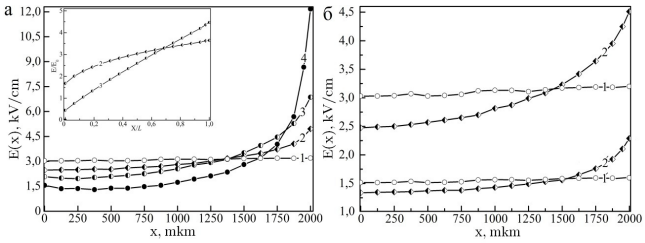


Fig. 4. *Steady-state distributions of electric field on MTISTIM CdZnTe structure: a - under the constant bias voltage V_0 = 601 V and illumination with light pulses of different power P: 1 - 0, 2 ~ 7.5 mW, 3 ~ 12.5 mW, 4 ~ 24.5 mW; the inset shows the coordinate dependences E(x) for the CdZnTe structure, calculated using the method [17] at the same values of the external macroparameters; b - under the constant pulsed pump power P_0 ~ 13 mW and different bias voltages V_0: 1 - 601 V, 1′ - 303 V, L ~ 2 mm, T = 300 K.*

At the same time, at high illumination intensities and fields E_0 ~ 4–6 kV/cm, the coordinate dependence $E(x)$ in the CdZnTe sample changes significantly: the region of the "strong" field is spatially localized in the narrow near-contact region of the non-illuminated electrode, occupying ~1/5–1/7 part of the sample, where it sharply increases with a positive curvature $d^2E/dx^2 > 0$ near the non-illuminated electrode and more weakly near the illuminated one, "sagging" inside the most part of the base. In this case, experiments show that in a number of samples studied, a slight increase in the field near the illuminated electrode may be absent, which is rather associated not with uncompensated impurity levels in the contact area of the illuminated electrode, but with asymmetric conditions for the passage though tunneling dielectric layers by carriers of different signs and the absence of injection from the contacts. In this case, the dependence $E(x)$ (curve 4 in Fig. 4*a*) can be approximated by the expression $E(x) \sim Ax^n$, where $n \sim 3.51$, which is in good agreement with the results of theoretical analysis, which in the diffusion-drift approximation predicts a superlinear dependence $E(x)$ in the region of monopolar transport, when the accumulation of charge in the near-contact region of the non-illuminated electrode cannot be neglected [8].

Another effect on the coordinate dependence $E(x)$ is exerted by a change in the bias voltage at a constant optical pump power. From the stationary distributions $E(x)$ measured in the same sample at different bias voltages V_0 and illumination with a fixed-power light pulse (Fig. 4*b*), it follows that a decrease (increase) in the magnitude of the applied voltage V_0 leads to a decrease (increase) in the value of the field E_0, by the magnitude of the changing voltage, while the coordinate field changes due to the monopolar mobile charge of the photocarriers change little. In other words, in such structures, the threshold field strength E_p is determined by the combination of the field E0 created by the external voltage source V_0 and the field E_1 created by the photoinduced space charge [8]:

$$E_p = E_0 + E_1. \quad (3)$$

It follows from Eq. (3) that the appearance of oscillation in such structures can be achieved with various relations between the values of external macro parameters. This feature leads to the fact that after the formation of the domain, changes in any of the macroparameters leads to a change in the frequency of the oscillations, which is observed in the experiment.

At the same time, the results of studies concerning the domain instabilities in high-resistance CdZnTe are not found in the literature. Therefore, along with the measurement of E_p in several CdZnTe samples, the velocity-field dependence $\upsilon(E)$ was measured. The measurement of the velocity of carriers under the conditions of inhomogeneous field $E = E(x)$ was carried out using the technique [9], which is based on scanning the sample from one electrode to another with a narrow beam of probe light that causes no photoactive absorption and measuring the transit time of carriers and the spatio-temporal and amplitude characteristics of the pulses of the probe light beam at each scanning step under illumination the sample with short light pulses under constant applied voltage. This approach allows calculating at each step the field value in the MDTPTDM structure and the carrier transit time corresponding to this value for any polarity of the applied voltage and any coordinate dependence $E(x)$.

Measurement of the dependence $\upsilon(E)$ in samples with different L showed that although the carrier velocity varies from sample to sample in the range υ ~0.7–1.25·10^7 cm/s, the maximum carrier velocity is reached at threshold fields E_p ~12.5–13.3 kV/

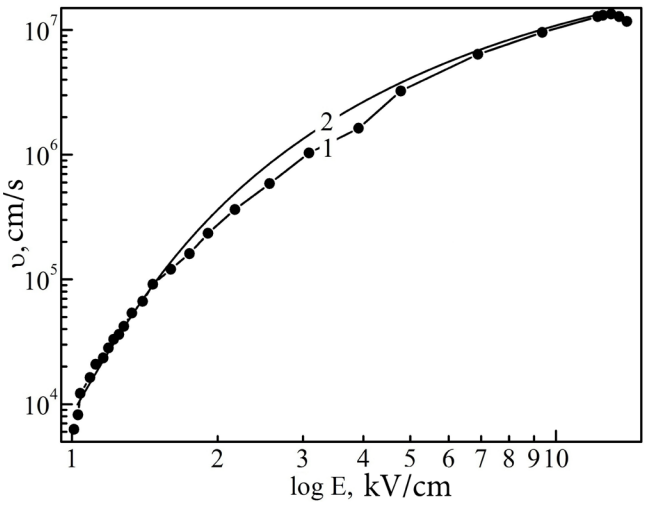


Fig. 5. *Dependence of the electron drift velocity υ on the electric field strength E(x) in CdZnTe: 1 - experiment, 2 - calculation; L = 2 mm, T = 300 K.*

cm, after which on the curve $\upsilon(E)$ a falling section is observed (**Fig. 5**).

The possibility of detecting the Gunn effect was theoretically analyzed earlier in [18], where, based on the results of Ref. [19], an expression is given that allows estimation of the carrier concentration necessary for the appearance of Gunn oscillations in a high-resistance semiconductor:

$$n > -\frac{2.09\varepsilon}{Lq\frac{1}{v}\left|\frac{dv}{dE}\right|}, \quad (4)$$

where E is the magnitude of the electric field strength, n is the concentration of electrons in the semiconductor before the beginning of domain formation, ε is the permittivity, q is the charge of electron, L is the separation between the contacts, υ is the absolute value of the electron drift velocity, $d\upsilon/dE$ is the differential mobility, which is negative due to the intervalley transitions of hot electrons.

It was also shown there that the generation due to electrical injection of carriers from a contact into a high-resistance n^+-n-n^+ structure is not possible, because at certain bias voltages V_0', the current mode limited by the spatial charge is replaced by the current mode limited by the emission ability of the cathode contact n^+-n, and the field strength remains constant regardless of the further growth of V_0'. However, in an illuminated high-resistance tunnel MTISTIM structure, these restrictions are removed, because after the establishment of the nonequilibrium depletion regime, the rate of generation of nonequilibrium photocarriers in such structures is proportional to the illumination intensity [20], and an increase in the photoinduced mobile charge leads to an increase in the field in one of the regions of the structure. As a result, due to nonequilibrium carriers Δn under illumination it becomes possible to provide the carrier concentration n and the field strength E_p necessary for the occurrence of oscillation. In addition, consequently, this leads to the fact that, firstly, the oscillation without illumination in such structures cannot occur at any bias voltages, and, secondly, due to this mechanism, an unambiguous relationship between the duration of oscillation and the duration of the pump light pulse.

Using the oscillograms of photocurrent pulses shown in Fig. 3a, one can estimate the initial carrier concentration n at which stable oscillation occurs in the CdZnTe structure. It should be noted that, taking Eq. (3) into account, the value of the threshold field E_p was determined at values of external macroparameters close to the threshold.

Then for a sample of CdZnTe, assuming for estimate $L = 4.61\cdot10^{-2}$ cm, ε = 10.3 (1 MHz) [21], $E_p \sim 1.28\cdot10^4$ V/cm, $q = 1.6\cdot10^{-19}$ C, $E_0 \sim 8.7$ kV/cm, $\upsilon \sim 9\cdot10^5$ cm/s, $\left|d\upsilon/dE\right| \sim 70.3$ $cm^2V^{-1}s^{-1}$, the initial concentration of carriers n amounts to $\sim 3.5\cdot10^{12}$ $см^{-3}$, and the parameter $nL \sim 1.61\cdot10^{11}$ cm^{-2} [4].

For the mean current photosensitivity $S_i \sim 0.75$ A/W and the reflection coefficient $R \sim 0.3$-0.35, the light power, at which the oscillation arises, is ~7–9 mW, which agrees well with the experimental data.

From the oscillograms (Fig. 3 b,c) it follows that changing any of the macroparameters within the range from P_p to P_{max} at $V_0 = V_p$ or from V_p to V_{max} at $P_0 = P_p$ results in the practically similar maximal frequency of oscillation fmax, increasing from $f_1 \sim 19.6$ MHz to $f_m \sim 260$ MHz under the growth of the optical pump power from P_p to P_{max} with the step $\Delta P \sim 1$ mW and the bias voltage from V_p to V_{max} with the step $\Delta V \sim 5.2$ V.

5. CONCLUSIONS

This paper presents the results of experimental studies of a new type of domain instability, which is based on drift nonlinearity – the photostimulated

spatial restructuring of the electric field. It is experimentally shown that in a high-resistance tunnel MIS structure, this physical mechanism leads to a significant deviation from equilibrium of its main macroparameter, the electric field, which becomes unstable at certain values of the macroparameters and jumps from the stationary nonuniform distribution throughout the entire MTISTIM structure to a new, but also steady state, in which it becomes narrowly localized and periodically moves from one electrode to another.

Accordingly, in MTISTIM structures based on electro-optical crystals, changing the electro-optical characteristics of the medium by the domain field makes it possible, due to the transverse electro-optical Pockels effect, to transfer optical information from the controlling light beam $I_1(x, t)$ to another light beam $I_2(x, t)$, which is transmitted through the structure, i.e. to carry out high-frequency optical modulation of one light flux by another.

REFERENCES

1. Bratchikov AN, Ioannesyants MR. Analiz sovremennogo sostoyaniya i tendentsiy razvitiya elementnoy bazy optovolokonnykh sistem antennykh reshetok [Analysis of the current state and development trends of the element base of fiber-optic systems of antenna arrays]. *Uspekhi sovremennoy radioelektroniki,* 1997, 7:3-15.
2. Vasil'yev PP. Pikosekundnaya optoelektronika. *Kvantovaya elektronika*, 1990, 17(3):268-287 [Picosecond optoelectronics. *Soviet Journal of Quantum Electronics*, 1990, 20(3):209–227].
3. Usanov DA, Skripal' AV. F*izika raboty poluprovodnikovykh priborov v skhemakh SVCH* [Physics of semiconductor devices in microwave circuits]. Saratov, SGU Publ., 1999, 376 p.
4. Vorobyov LE, Danilov CH, Ivchenko EL, Levinstein ME, Firsov DA, Shalygin VA. *Kineticheskiye i opticheskiye yavleniya v sil'nykh elektricheskikh polyakh v poluprovodnikakh i nanostrukturakh* [Kinetic and optical phenomena in strong electric fields in semiconductors and nanostructures]. St. Petersburg, Nauka Publ., 2000, 160 p.
5. Levinshteyn ME. Vliyaniye osveshcheniya na parametry diodov Ganna [The effect of lighting on the parameters of Gunn diodes]. *Fizika i tekhnika poluprovodnikov*, 1973, 7(7):1332-1337 (in Russ.).
6. Gontar VM, Yegnazaryan GA, Rubin VS, Murygin VM, Stafeev VI. Otritsatel'naya differentsial'naya provodimost' v vysokoomnom arsenide galliya pri osveshchenii [Negative differential conductivity in high-resistance gallium arsenide under illumination]. *Fizika i tekhnika poluprovodnikov,* 1971, 5(6):1061-1066 (in Russ.).
7. Bonch-Bruyevich VL, Kalashnikov SG. *Fizika poluprovodnikov* [Semiconductor Physics]. Moscow, Nauka Publ., 1990, 686 p..
8. Kasherininov PG, Kichayev AV, Perepelitsyn YuN, Khartsiyev VE, Yaroshetskiy ID. Fotoelektricheskiye yavleniya v poluprovodnikovykh strukturakh s fotochuvstvitel'nym raspredeleniyem elektricheskogo polya i optoelektronnyye pribory na ikh osnove. Chast' I [Photoelectric phenomena in semiconductor structures with a photosensitive distribution of the electric field and optoelectronic devices based on them. Part I]. St. Petersburg, *Preprint FTI-1569*, 1991, 59 p.
9. Zhavoronkov NV, Perepelitsyn YuN. Issledovaniye vozmozhnosti sozdaniya fotochuvstvitel'nogo elementa dlya dvumernykh zerkal fazovrashchateley i akustoopticheskikh perestraivayemykh fil'trov v opticheski upravlyayemykh fazirovannykh antennykh reshetkakh, funktsioniruyushchikh v IK diapazone spektra. [Investigation of the possibility of creating a photosensitive element for two-dimensional mirrors of phase shifters and acousto-optic tunable filters in optically controlled phased antenna arrays operating in the infrared range of the spectrum.] *Otchet po GK № 12411.1006899.11.102*. Moscow, Zelenograd, ZAO NIIMV Publ., 2012, 146 s.
10. Kosyachenko LA, Sklyarchuk VM, Sklyarchuk OV, Maslyanchuk OL Shirina zapreshchennoy zony kristallov CdTe i $Cd_{09}Zn_{01}Te$ [Band gap of CdTe and $Cd_{09}Zn_{01}Te$ crystals], *Fizika i tekhnika poluprovodnikov,* 2011, 45(10):1323-1330. [Band gap of CdTe and $Cd_{09}Zn_{01}Te$ crystals. *Semiconductors*, 2011, 45(10):1273-1280].

11. Perepelitsyn YuN, Zhavoronkov NV. Fotonnyye ustroystva obrabotki i kommutatsii opticheskogo signala [Photonic devices for processing and switching an optical signal]. *Mater. 1 Rossiysko-Belorusskoy NTK "Elementnaya baza otechestvennoy radioelektroniki", 1:132-135 (N. Novgorod, 2013).* N.Novgorod, NGU Publ., 2013.

12. Perepelitsyn YuN, Zhavoronkov NV. Odnokanal'nyye ustroystva obrabotki opticheskogo signala. [Single-cGunnel optical signal processing devices]. *Foton-ekspress*, 2016, 6:202-203.

13. Lampert M, Mark P. *Current injection in solids.* New York: Academic Press, 1970, 351 p.

14. Kasherininov PG, Kichayev AV, Yaroshetskiy ID. Raspredeleniye napryazhennosti elektricheskogo polya v vysokoomnykh M(TD)P(TD)M strukturakh pri osveshchenii [Electric field strength distribution in high-resistance M(TI) S(TI)M structures under illumination]. *Pis'ma v zhurnal tekhnicheskoy fiziki,* 1993, 19(17):49-54.

15. Kasherininov PG, Kichayev AV, Yaroshetskiy ID. Fotoelektricheskiye yavleniya v vysokoomnykh strukturakh s granitsey razdela poluprovodnik – tonkiy sloy dielektrika na vysokoomnykh kompensirovannykh kristallakh [Photoelectric phenomena in high-resistance structures with a semiconductor-thin insulator layer interface on a high-resistance compensated crystal]. *Zhurnal tekhnicheskoy fiziki,* 1995, 65(9):193-196.

16. Kasherininov PG, Kichayev AV, Tomasov AA. Fotoelektricheskiye yavleniya v strukturakh na vysokoomnykh poluprovodnikovykh kristallakh s tonkim sloyem dielektrika na granitse poluprovodnik-metall [Photoelectric phenomena in structures based on high-resistance semiconductor crystals with a thin dielectric layer at the semiconductor-metal interface]. *Fizika i tekhnika poluprovodnikov,* 1995, 29(11):2092-2107 (in Russ.).

17. Kasherininov PG, Reznikov BI, Tsarenkov GV. Fotoeffekt v strukturakh metall – poluprovodnik – metall na osnove vysokoomnogo poluprovodnika [Photoeffect in a metal-semiconductor-metal structure made of a high-resistivity semiconductor. *Sov. Phys. Semicond.* , 1992, 26, 832.

18. Ryabinkin YUS. O popytke obnaruzheniya effekta Ganna v rezhime prostranstvennogo zaryada [On an attempt to detect the Gunn effect in space charge mode]. *Fizika i tekhnika poluprovodnikov,* 1963, 2(8):1168-1170 (in Russ.).

19. McCumber DE, Chynoweth AG. Theory of negative conductance amplification and of Gunn-instabilities in "two-valley" semiconductors. *IEEE Trans. Electron. Devices*, ED-13(4):4-21.

20. Vul' AYa, Dideykin AT, Kozyrev SV. Fotopriyemniki na osnove struktur metall–dielektrik–poluprovodnik [Photodetectors based on metal-dielectric-semiconductor structures]. *Fotopriyemniki i fotopreobrazovateli.* Leningrad, Nauka Publ., 1986, p. 105-107.

21. Yu P, Cardona M. *Osnovy fiziki poluprovodnikov.* Berlin-Heidelberg: Springer-Verlag, 2010, 778 p.

DOI: 10.17725/rensit.2020.12.201

On the Design of Rectenna

Mhnd Farhan

University of Baghdad, http://www.uobaghdad.edu.iq/
Baghdad, Iraq

E-mail: mhndfarhan@yahoo.com

Received March 04, 2020; peer reviewed March 16, 2020; accepted March 30, 2020

Abstract. **This paper focuses on designing a circuit that rectifies background radiation and one that is self-biasing. This circuit set-up is called a rectenna which is a special type of antenna that is used to convert radio-frequency energy into direct current electricity. A simple model of a rectenna element consists of a monopole antenna with an radio frequency(RF) diode bridge connected in series with the antenna. The bridge rectifies the ac current induced in the antenna by the electromagnetic radiation to produce dc power which is used to bias a Bipolar Junction transistor(BJT). RF sensitive/high switching diodes are usually used because they have the lowest voltage drop and highest speed and therefore have the lowest power losses due to conduction and switching. The BJT transistor has a feedback biasing and essentially amplifies the ac signal from the antenna. The amplified signal is fed into an RF diode for dc conversion. There are two stages of amplification in order to achieve a big voltage magnitude at the output that can be used to charge a device with low power ratings. Thus the idea of a cell-less power source is achieved in such implementation.**

Keywords: **rectenna; cell-less power source; design**

UDC 621.396.6

For citation: Mhnd Farhan. On the Design of Rectenna. *RENSIT,* 2020, 12(2):201-206; DOI: 10.17725/rensit.2020.12.201.

Contents

1. INTRODUCTION

Background radiation is electromagnetic radiation due to mobile telephone systems and broadcasting in transmission of information. This electromagnetic energy can be converted to electrical energy and used to power electrical devices.

Conversion to electrical energy is effected by use of a rectenna. A rectenna or a rectifying antenna is a special type of antenna that is used to convert radio frequency (RF) energy into direct current electricity. They are used in wireless power transmission systems that transmit power by radio waves. A simple rectenna element consists of a monopole/dipole antenna with an RF diode connected across the monopole/dipole elements. The diode rectifies the ac current induced in the antenna by the microwaves, to produce dc power, which powers a load connected across the diode. Large rectennas consist of an array of many such dipole elements [1,2].

The research in rectennas is diversified whereby different types of rectennas perform differently to achieve the same objectives. These include RF rectenna and optical rectennas.

With optical rectennas, it has been theorized that similar devices, scaled down to the proportions used in nanotechnology, could be used to convert light into electricity at greater efficiencies than

what is currently possible with solar cells.

Theoretically, high efficiencies can be maintained as the device shrinks, but experiments have so far only obtained roughly 1% efficiency while using infrared light.

Essentially, the use of batteries as energy source can be eliminated by using wireless powering. Besides recharging an integral battery is no more necessary for continuous telemetry operation. Depending on the power consumption requirement of a device or system, wireless remote powering can be performed with either near-field inductive coupling or far-field electromagnetic coupling. The choice of the remote powering frequency is based on the constraints of the application such as power consumption, device size, read range or proximity, transmission medium and data rate. This improvised system of device charging is what is ideally referred to as a cell-less power source. While a high data rate and high read range make it necessary to use high frequency communication, higher power delivery makes the use of near field powering more preferable. However the methodology in this paper implements far field powering [3,4].

A method to justify the above problem is the design of an antenna for its target integrated rectifier for far-field electromagnetic energy. Harvesting is done by use of RF antenna for pick-up power where the energy signal from the electromagnetic radiation is fed into the rectifier.

However, unavoidable process variations cause different input impedances and efficiency performances depending on the process corner of the fabricated chip.

The above methodology can be implemented with an addition of npn transistor and high switching diode. This enables amplification of the input ac signal from the antenna and the amplified signal is converted to dc by the diode to obtain a reasonable output. Therefore dual stage amplification is necessary to obtain a bigger magnitude output[5,6].

2. SYSTEM DESIGN

The most suitable choice for implementation is the design of a monopole antenna used with a rectifier built with well-characterized commercially available rectifier. This enables us to evaluate the performance of the rectenna more accurately, in order to achieve a reasonable overall efficiency and a good conversion efficiency at the incident frequency.

This rectenna uses monopole antenna principle, which feeds an RF diode quad bridge. The bridge is capable of handling high frequencies where FR207 diodes (RF sensitive/high switching) are used.

This rectenna element operates efficiently at much lower incident power levels of 20–65 mW with little reflected RF power (i.e., the overall efficiency is 1% lower than the conversion efficiency).This characteristic has two important applications in microwave-power beaming systems:

1. Power can be converted efficiently at the edge of the rectenna where power densities are lower than the center elements.
2. Power can be converted efficiently when the transmission distance is large and power density is low

2.1 Theory of operation

The circuit shown in **Fig. 1** is modified and additional components added whereby a monopole antenna attaches to a quad bridge, which transforms the monopole impedance to

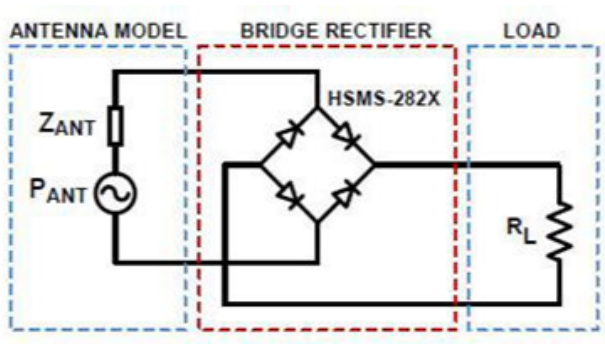


Fig. 1. *The main components of the rectenna element.*

the bridge impedance and rejects higher order diode harmonics from radiating through the monopole.

In order to ensure that the source and rectifier impedance are matched for obtaining correct input impedance of the rectifier, we use the built-in optimizer of Advanced Design System(ADS). It minimizes the reflection coefficient between the source and the load which therefore maximizes by finding conjugate input impedance with the given input power. **Fig. 2** shows simulation results of the input impedance of the rectifier versus frequency for 6 dBm of input power. Each point of the curve corresponds to matched source impedance as determined by the optimizer.

Suppose that the target input power is 6 dBm for the selected rectifier. Assuming the rectifier impedance ZRECT = 11.7−108 Ω by calculation. Maximum power transmission condition states that the rectifier and antenna have conjugately matched impedances. Therefore the target antenna impedance is determined as ZANT = 11.7 + 108 Ω.

The maximum simulated gain of the antenna is found as –5.1 dB and input impedance of the antenna is found as ZANT= 11.7 + 108 Ω. The antenna having dimensions of 12mm×10mm with dielectric thickness of 0.5 mm is fabricated on Rogers RO4003C dielectric with (ε_r = 3.55).

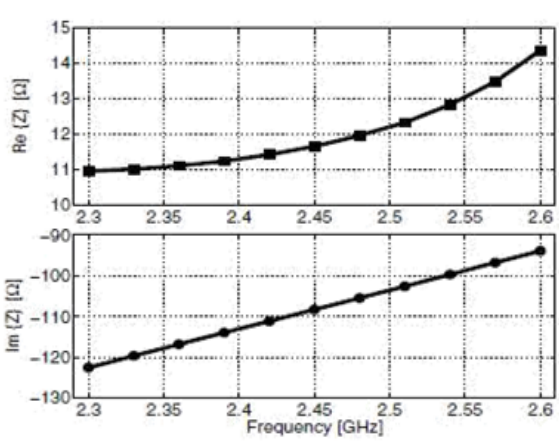


Fig. 2. *Simulation results of the input impedance of the rectifier.*

The typical operation of a rectenna element can be better understood by analyzing the bridge's dc characteristics with an impressed RF signal.

This simple model as illustrated in Fig.1 assumes that the harmonic impedances seen by the diode are either infinite or zero to avoid power loss by the harmonics. Thus, the fundamental voltage wave is not corrupted by higher order harmonic components.

The rectenna conversion efficiency then depends only on the diode (bridge) electrical parameters and the circuit losses at the fundamental frequency and dc.

Fig. 3 shows the equivalent circuit of the diode used for the derivation of the mathematical model. The diode parasitic reactive elements are not included in the equivalent circuit. Instead, it is assumed they belong to the rectenna's environment circuit.

The environment circuit is defined as the circuit around the diode that consists of linear-circuit elements.

The diode model consists of a series resistance R_s, a nonlinear junction resistance R_j described by its dc IV characteristics, and a nonlinear junction capacitance C_j. A dc load resistor is connected in parallel to the diode along a dc path represented by a dotted line to complete the dc circuit. The junction resistance R_j is assumed to be zero for forward bias and infinite for reverse bias.

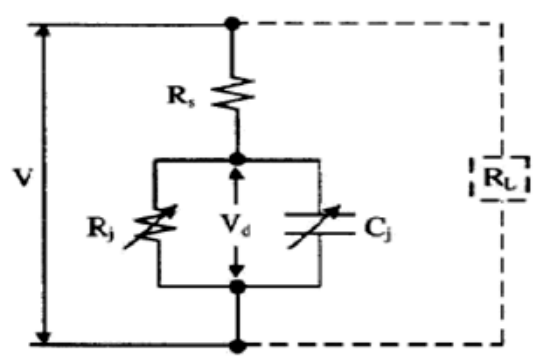


Fig. 3. *Equivalent circuit model of the rectifying circuit.*

2.2 Actual design

Fig. 4 is an illustration of a self-biasing circuit adopted from the simulation and a graph demonstrating a dc output voltage.

With the incident input power ranging from 20-65 mw, the antenna produces approximately 4.03 v and current of 10.7 mA which is got from the relation $P = V^2/R$ where R is the intrinsic impedance (120 π) with incident power of 43 mw (16.33 dBm). The antenna in the circuit diagram is represented by a power source.

The monopole antenna is adjusted accordingly to give the maximum pick-up power. With a pick-up test done using a radio, it is observed that the best operating frequency is at 100.3 MHz. From the relation $c = f\lambda$, the wavelength is calculated to be 3 m. From the monopole antenna characteristics, the wavelength is $\lambda/4$ which is approximately 0.75 m (75 mm) and thus the antenna is adjusted to this length to give maximum efficiency.

At the rectifier stage, RF sensitive diodes (FR207) are used which are sensitive to high frequencies ranging from 85 MHz to 110 MHz and thus ac signal from the antenna is rectified to dc signal and obtained at the output of the bridge. This dc signal is just sufficient to bias a bipolar junction npn transistor to its quiescent point.

The transistor is set-up in the feed-back/ self-biasing configuration. This self-biasing configuration is another Beta(β) dependent biasing method that requires only two resistors to bias the transistor. The collector to base feedback configuration ensures that the transistor is always biased in the active region regardless of the value of Beta (β) as the base bias is derived from the collector voltage.

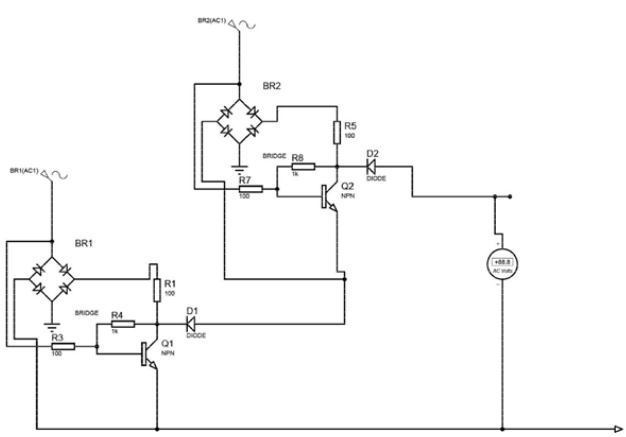


Fig. 4. *Circuit diagram of the design implementation from simulation.*

In this circuit, the base bias resistor (RB) is connected to the transistors collector, instead of to the supply voltage rail (Vcc). Now if the collector current increases, the collector voltage drops, reducing the base drive and thereby automatically reducing the collector current. Then this method of biasing produces negative feedback.

The biasing voltage is derived from the voltage drop across the load resistor (RL). So if the load current increases there will be a larger voltage drop across RL, and a corresponding reduced collector voltage (VC) which will cause a corresponding drop in the base current (IB) which in turn, brings (IC) back to normal.

The opposite reaction will also occur when transistors collector current becomes less. Then this method of biasing is called self-biasing with the transistors stability using this type of feedback bias network being generally good for most amplifier designs.

With this configuration, the transistor essentially amplifies the ac signal from the antenna and has a feedback factor/gain of 10. The feedback path also provides the system with stability. With the amplified output obtained from the transistor element; it passes through a diode to convert to dc.

There are two stages of amplification in the above setup, each with equal parameters set whereby the output of the first stage is fed into the second stage for further amplification. The combined output magnitude of the two stages is further amplified using an operational amplifier with a feedback factor/gain of 10 to achieve a

reasonable magnitude of 5.3 V that can be fed into a device charging systems rated at 5 V-10 V. When a 370 Ω load is put across the output terminals, a reasonable current of 200 mA is achieved at the output of the operational amplifier.

3. RESULTS AND ANALYSIS

3.1. Incident power and voltage analysis

With the input power range already known to be 20-65 mW, a relation of incident antenna power in free space to voltage is established. A graph of varying input power level density in mill-watts with voltage (V) at intrinsic impedance is plotted as shown in **Fig. 5**.

The above relationship demonstrates that voltage input increases with increasing power density which implies that the antenna should be adjusted such that it receives optimal power and thus obtain maximum conversion efficiency. The differences between the overall efficiency and conversion efficiency indicate that little power (< 1%) is reflected.

3.2. Input and output voltage analysis

With the varying input voltages, a relation is drawn with the output of the dual amplification level at 1millisecond of simulation as shown in **Fig. 6**.

The above behavior is observed demonstrating some erraticism and showing that the output power levels vary with minor disparities. These minor disparities occur at the expected input voltage range (2.7-5) V. The disparities could be credited to the fact there are losses in the input power due to mismatch between the rectifier and the antenna. Assuming minimal mismatch losses in the rectenna design, the conversion efficiency of the diode is limited primarily by the rectifying diode. As such recommendations towards determining ways of increasing the conversion efficiency of the diode other than through the change of load resistance is beneficial.

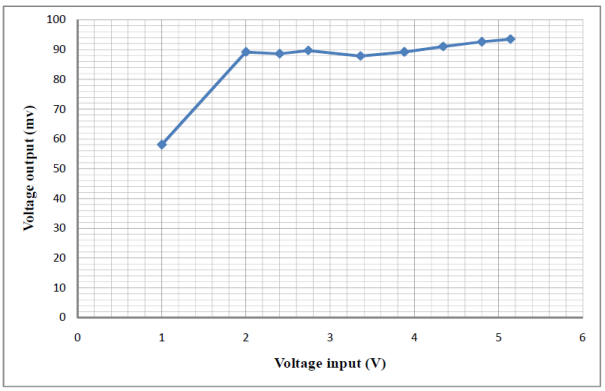


Fig. 6. *Input and output voltage analysis.*

3.3. Voltage-current characteristics

Fig. 7 demonstrates ohms law showing that the simulation is accurate with the results obtained when a load of 200 Ω is connected across the output terminals.

However this observation implies that with a fixed load at the output, the variation of electromagnetic radiation could alter the current value at the output causing a spike or a value which is lower than the rating. With the use of a

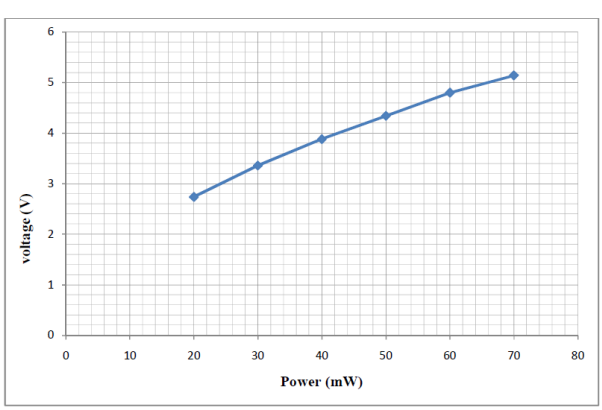


Fig. 5. *Incident antenna power versus voltage.*

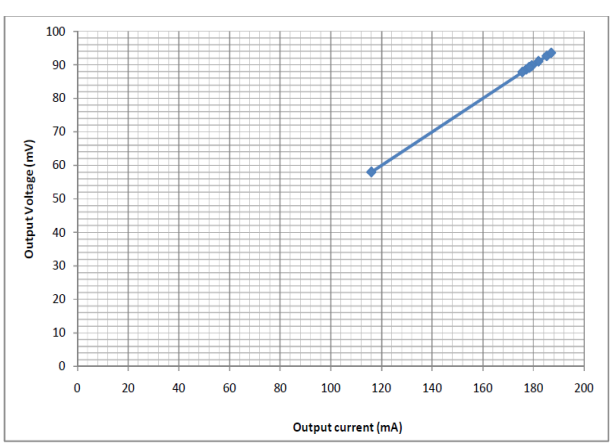


Fig. 7. *Voltage-current characteristics.*

regulator, the output can be stabilized although this doesn't control the input power density.

The breakdown voltage of the diode limits the power handling capability of each rectifying circuit. The diode model needs to be valid for a wide range of biasing in order not to affect the output voltage. This will in turn ensure the current at the load is not affected and the system therefore operates successfully.

4. CONCLUSION

The work in this paper was mainly implemented through a run simulation and the results obtained were reasonable in contrast with practical realization of previously done implementations.. The objectives of the experiment were achieved whereby self-biasing circuit was created and produced output that can be used to charge a device with a rating of at least 5 V. The antenna allows 100 MHz operating frequency to pass and diodes used prevent interference of signals and re-radiation of higher order harmonics generated due to the RF sensitivity and high switching property.

REFERENCES

1. J. Shin, M. Seo, and J. Choi. A compact and wideband circularly polarized rectenna with high efficiency at X-band. *Progress In Electromagnetics Research*, 2014, 145:163-173.
2. C. Song, Y. Huang, J. Zhang, and S. Yuan. A high-efficiency broadband rectenna for ambient wireless energy harvesting. IEEE *Transactions on Antennas and Propagation*, 2015, 63(8):3486-3495.
3. H. Mei, X. Yang, B. Han, and G. Tan. High-efficiency microstrip rectenna for microwave power transmission at Ka band with low cost. *IET Microwaves Antennas & Propagation*, 2016, 10(15):1648-1655.
4. H. Sun, and G. Wen. A new rectenna using beamwidth-enhanced antenna array for RF power harvesting applications. *IEEE Antennas & Wireless Propagation Letters,* 2016, 16:1451-1454.
5. D. Gretskih, A.V. Gomozov, V.A. Katrich. Mathematical model of large rectenna arrays for wireless energy transfer. *Progress In Electromagnetics Research,* 2017, 74:77-91.
6. Y. Shi. Design of a novel compact and efficient rectenna for WiFi energy harvesting. *Progress In Electromagnetics Research C*, 2018, 83:57-70.

DOI: 10.17725/rensit.2020.12.207

A Survey of Software Radios: Reconfigurable Platforms, Development Tools, and Future Directions

[1]Hassan Nasser, [1]Abdelrazak Badawieh, [2]Abdulkarim Assalem

[1]Damascus University, http://damascusuniversity.edu.sy/
Damascus, Syrian Arab Republic
[2]Al-Baath University, http://albaath-univ.edu.sy/
Homs - p. 77, Syrian Arab Republic

E-mail: hsnsy.aa@gmail.com, Gbadawil@gmail.com, assalem1@gmail.com
Received March 11, 2020; peer reviewed March 20, 2020; accepted March 27, 2020

Abstract. **Software-Defined Radio (SDR) approaches for rapid prototyping of radio systems using reconfigurable hardware platforms offer significant advantages over traditional analog and hardware-centered methods. In particular, time and cost savings can be achieved by reusing tested design artefacts; this translates to supporting various features and functionalities, such as updating and upgrading through reprogramming, without the need to replace the hardware on which they are implemented. This opens the doors to the possibility of realizing multi-band and multi-functional wireless devices. Progress in the SDR field has led to the escalation of protocol development and a wide spectrum of applications, with more emphasis on programmability, flexibility, portability, and energy efficiency, in Mobile technology, Wi-Fi, and M2M communication. Consequently, SDR has earned a lot of attention and great significance to both academia and industry. SDR designers intend to simplify the realization of communication protocols while enabling researchers to experiment with prototypes on deployed networks. This Research is a survey of the state-of-the art SDR platforms and development tools in the context of wireless communication which presented an overview of SDR architecture and its basic components; and discussed the significant design trends and development tools. In addition, we reviewed available SDR platforms with an analytical comparison based on a set of metrics as a guide to developers. Finally, we offered some predictable future Directions for SDR Researches**

Keywords: **Software Defined Radio, Reconfigurability, FPGAs, Platforms, System on Chip, Radio Communications**

PACS: 01.20.+×

For citation: Hassan Nasser, Abdelrazak Badawieh, Abdulkarim Assalem. A Survey of Software Radios: Reconfigurable Platforms, Development Tools, and Future Directions. *RENSIT,* 2020, 12(2):207-218; DOI: 10.17725/rensit.2020.12.207.

Contents

1. INTRODUCTION

According to Cisco Systems, 500 Billion wireless devices for 7 Billion people are expected to be connected to the internet by 2030 [1]. Each device includes sensors that collect data, interact with environment, and communicate over a network. Therefore, the first challenge is to adjust the basic connectivity and networking layers to handle the large numbers of ends [2, 3]. There is an increasing number of wireless protocols that have been developed, such as BLE, LTE, and new Wi-Fi protocols for Machine-to-Machine (M2M) communication purposes due to different demanding requirements, one of which is high-energy efficiency. Wireless standards, generally,

are adapting quickly in order to accommodate different user needs and hardware specifications. This has called for a transceiver design with the ability to handle several protocols, including the existing ones and those being developed. In order to accomplish this task, one needs to realize the protocols need for a flexible, reconfigurable, and programmable framework.

THE National Aeronautics and Space Administration (NASA) is developing an on-orbit, adaptable, Software Defined Radio (SDR)/Space Telecommunications Radio System (STRS)-based testbed facility to conduct a suite of experiments to advance technologies, reduce risk, and enable future mission capabilities on the International Space Station (ISS). The Space Communications and Navigation (SCaN) Testbed Project will provide NASA, industry, other Government agencies, and academic partners the opportunity to develop and field communications, navigation, and networking technologies in the laboratory and space environment based on reconfigurable, SDR platforms and the STRS Architecture. Led by the NASA Glenn Research Center (GRC), the SCaN Testbed was developed to move SDR space technology forward, as the advancements of Appendix Table 2 indicate. The project was previously known as the Communications, Navigation, and Networking reconfigurable Testbed (CoNNeCT) [5].

Global Industry Analysts highlights some of the market trends for SDR as follows [4]: (a) increasing interest from the military sector in building communication systems and large-scale deployment in developing countries, (b) growing demand for public safety and disaster preparedness applications, and (c) building virtualized base stations (BSs). SDRs are also ideal for developing future space communications [6], Global Navigation Satellite System (GNSS) sensors [7], Vehicle-to-Vehicle (V2V) communication [8, 9], and Internet of Things applications [10, 11], where relatively small and low-power SDRs can be utilized.

SDRs are implemented through employing various types of hardware platforms, such as General Purpose Processors (GPPs), Graphics Processing Units (GPUs), Digital Signal Processors (DSPs), and Field Programmable Gate Arrays (FPGAs). Each of these platforms is associated with their own set of challenges. Some of these challenges are: (a) Utilizing the computational power of the selected hardware platform, (b) keeping the power consumption at a minimum, ease of design process, (c) Cost of tools and equipment. The research community and industry have both developed SDRs that are based on the aforementioned hardware platforms.

2. RESEARCH METHODOLOGY AND GOALS

In this Research, we first introduced some principles and criterions that wireless communications based on. Then we presented an overview of SDR architecture as well as the analog and digital divides of the system and interconnection of components. Furthermore, we reviewed the SDR platforms developed by both industry and academia, and provided an analytical comparison of hardware platforms as a guide for design decision making. Moreover, we discuss the use of respective development tools and present a summary to help explain their functionalities and the platforms they support.

This Research is organized as follows: Section 3 provides a description of SDR architecture. Section 4 Design Mythologies & Platforms and comparison of the commercially and academically developed SDR platforms. Section 5 we review the common development tools that are typically used in the process of SDR design and implementation for different design approaches. Finally, we offered some predictable future directions for SDR researches, and concluded the paper in Section 6.

3. PRINCIPLES AND ARCHITECTURE

First, we will overview some basic principles:

- **Table 1** shows the Electromagnetic Spectrum Based on the International Telecommunications Union (ITU) [12]:
- The term **Radio Spectrum Policy** generally refers to "electromagnetic frequencies between 9 KHz and 3000 GHz with wavelengths between one millimeter and thousands of kilometers".
- In **conventional Radio**: Radio Components are implemented as analog parts or static silicon, while in *Software Radio* reconfigurable parts are used instead; and make the components generic so they could be used to implement several types of Radios. **Figure 1** shows this idea:
- The Institute of Electrical and Electronic Engineers (IEEE) P1900.1 Working Group [13] has created the following definition for the

Table 1

Electromagnetic Spectrum and Wavelength types

Electromagnetic spectrum						
Gamma rays	X-rays	Ultraviolet	Visible	Infrared	Microwave	Radio
ν ZHz	EHz	PHz	THz	GHz	MHz	
fm	pm	nm	μm	mm	m	km λ
← higher frequencies			longer wavelengths →			

X-rays	Ultraviolet	Visible (optical)	Microwaves	Radio spectrum (ITU)		Wavelength types
• soft X-ray • hard X-ray	• Extreme ultraviolet • Vacuum ultraviolet • Lyman-alpha • FUV • MUV • NUV • UVC • UVB • UVA	• Violet • Blue • Cyan • Green • Yellow • Orange • Red	• W band • V band • Q band • Ka band • K band • Ku band • X band • C band • S band • L band	THF	300 GHz/1 mm 3 THz/0.1 mm	• Microwave • Shortwave • Medium wave • Longwave
				EHF	30 GHz/10 mm 300 GHz/1 mm	
				SHF	3 GHz/100 mm 30 GHz/10 mm	
				UHF	300 MHz/1 m 3 GHz/100 mm	
				VHF	30 MHz/10 m 300 MHz/1 m	
				HF	3 MHz/100 m 30 MHz/10 m	
				MF	300 kHz/1 km 3 MHz/100 m	
				LF	30 kHz/10 km 300 kHz/1 km	
				VLF	3 kHz/100 km 30 kHz/10 km	
				ULF	300 Hz/1 Mm 3 kHz/100 km	
				SLF	30 Hz/10 Mm 300 Hz/1 Mm	
				ELF	3 Hz/100 Mm 30 Hz/10 Mm	

Software-Defined Radio (SDR): "A Radio in which some or all of the physical layer functions are software defined".

As shown in **Figure 2**, at a high level, a typical SDR transceiver consists of the following components: Signal Processing, Digital Front End, Analog RF Front End, and an antenna.

1. Antenna: SDR platforms usually employ several antennas to cover a wide range of frequency bands [15]. Antennas are generally referred to as "intelligent or smart" due to their ability to select a frequency band and adapt with mobile tracking or interference cancellation [14]. In the case of SDRs, an antenna needs to meet a certain list of requirements such as self-adaptation (flexibility to tuning to several bands), self-alignment (beamforming capability), and self-healing (interference rejection) [16].

2. RF Front End: This is a RF circuitry that its

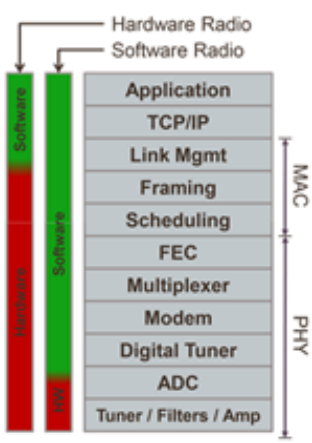


Fig. 1. *Hardware Radio vs Software Radio.*

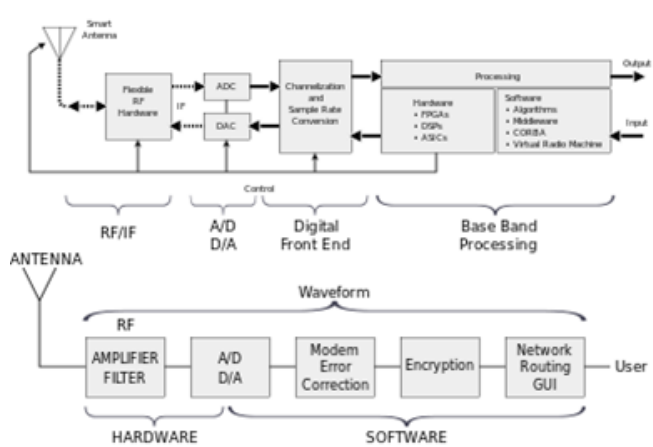


Fig. 2. *SDR transceiver architecture.*

main function is to transmit and receive the signal at various operating frequencies. Its other function is to change the signal to/from the Intermediate Frequency (IF). The process of operation is divided into two, depending on the direction of the signal (Tx or Rx mode):

- In the transmission path, the Digital-to-Analog Converter (DAC), which in turn feeds the RF Front-End, converts digital samples into an analog signal. This analog signal is mixed with a preset RF frequency, modulated, and then transmitted.
- In the receiving path, the antenna captures the RF signal. The antenna input is connected to the RF Front End using a matching circuitry to guarantee an optimum signal power transfer. It then passes through a Low Noise Amplifier (LNA), which resides in a close proximity to the antenna, to amplify weak signals and minimize the noise level. This amplified signal, with a signal from the Local Oscillator (LO), are fed into the mixer in order to down convert it to the IF [17].

3. Analog-to-Digital and Digital-to-Analog Conversion: The DAC, as mentioned in the previous section, is responsible for producing the analog signal to be transmitted from the digital samples. On the receiver side, the ADC resides, and is an essential component in radio receivers. The ADC is responsible for converting continuous-time signal to a discrete-time, binary-coded signals. ADC performance can be described by various parameters [18] including: **(1)** Signal-to-Noise Ratio (SNR): the ratio of signal power to noise power in the output, **(2)** resolution: number of bits per sample, **(3)** Spurious-free Dynamic Range (SFDR): the strength ratio of the carrier signal to the next strongest noise component or spur, and (4) power dissipation. Advances in SDR development have provided momentum for ADC performance improvements. For example, since ADC's power consumption affects the lifetime of battery-powered SDRs, more energy efficient ADCs have been developed [20].

4. Digital Front End: The Digital Front End has two main functions [19]:

- Sample Rate Conversion (SRC), which is a functionality to convert the sampling from one rate to another. This is necessary since the two communication parties must be synchronize.
- Channelization, which includes up/down conversion in the transmitter and receiver side, respectively. It also includes channel filtering, where channels that are divide by frequency are extracted. include interpolation and low-pass filters.

In a SDR transceiver, the following tasks are executed in the digital front end:

- In the transmitting side, the Digital Up Converter (DUC) translates the aforementioned baseband signal to IF. The DAC that is connected to the DUC then converts the digital IF samples into an analog IF signal. Afterwards, the RF up-converter converts the analog IF signal to RF frequencies.
- In the receiving side, the ADC converts the IF signal into digital samples. These samples are subsequently fed into the next block, which is the Digital Down Converter (DDC). The DDC includes a digital mixer and a numerically controlled oscillator. DDC extracts the baseband digital signal from ADC, and after being processed by the Digital Front End, this digital baseband signal is forwarded to a high-speed digital signal-processing block [21].

5. Signal Processing: The signal-processing block is referred to as the baseband processing block. When discussing SDRs, the baseband block is at the heart of the discussion, because it makes up the corpus of the digital domain of the implementation. This implementation runs on top of a hardware circuitry that is capable of processing signals efficiently as ASICs, FPGAs, DSPs, GPPs, and GPUs.

4. DESIGN MYTHOLOGIES & PLATFORMS

4.1. Field Programmable Gate Array-based:

1. Zu7/5/4 Zynq UltraScale+ MPSoC development kit.

2. Arria 10 SoC/FPGA FMC+ Development platform.

3. Zynq 7000 SODIMM Development kit.

4. Altera Cyclone V SoC Development Platform.

5. Airblue [35]: It is a method to implement radios on FPGA to achieve configurability by using an HDL language called Bluespec, through which all hardware blocks of a radio transceiver are written. In Bluespec, a developer describes the execution semantics of the design through Term Rewriting

Systems (TRS). TRS is a computational paradigm based on the repeated application of simplification rules [22]. The next step is compiling the TRS into RTL codes. TRS has the capability of generating efficient hardware designs. The main difference between a Verilog interface and a Bluespec interface is that the former is merely a collection of wires with no semantic meaning, while the latter includes handshake signals for blocks communication. Therefore, Bluespec facilitates latency insensitive designs, which are essential to system construction via modular composition. Using Airblue, developers may find the need to modify the building blocks, or modules, to add new features, make algorithmic modifications, and tune the performance to meet throughput or timing requirements, or make FPGA-specific optimizations.

4.2. Digital Signal Processor-based

1. Imagine Processor-based SDR: One of the earliest SDR solutions that is fully based on a DSP developed at Stanford University in 2001 [23]. The Stanford Imagine project aimed at providing a signal and image processor that was C programmable and was able to match the high performance and density of an ASIC. It is based on stream processing [24], which is similar to dataflow programming in exploiting data parallelism and is suitable for signal processing applications. This work paved the way to the development of GPUs.

2. University of Michigan's **Software On-Demand Architecture (SODA)**: A high performance SDR platform based on multi-core DSPs with one ARM Cortex-M3 processor for control purposes and multiple processing elements for DSP operations. Using four processing elements can meet the computational requirements of 802.11a and W-CDMA.

3. ARM Ardbeg: A commercial prototype based on SODA architecture. The main enhancements of Ardbeg compared to SODA are optimized SIMD design, Very Long Instruction Word (VLIW) support, and a few special ASIC accelerators, which are dedicated to certain algorithms such as Turbo Encoder/Decoder, Floating Point and Arithmetic Operations.

4. Atomix [34]: It is a declarative language describe the software in blocks, named atoms. We can implement any operation (signal processing or system handling) by an atom. Atoms can be used for realizing blocks, flow graphs, and states in wireless stacks. In addition, simple control flow makes atoms composability. It is important to note that an Atomix signal processing block implements a fixed algorithmic function, operates on fixed data lengths, is associated with a specific processor type, and uses only the memory buffers passed to it during invocation. The blocks will run fixed sets of instructions executing uninterrupted on fixed resources using fixed memories. This results in having fixed execution times. Atoms can also be composed to build larger atoms. Using Atomix, radio software can be built entirely out of atoms and is easily modifiable. Atomix based radio also meets throughput and latency requirements. The want of Atomix is that it is intended only for synthesis on a variety of DSPs, but not for GPPs, GPUs, or FPGAs.

5. BeagleBoard-X15: A cooperative project between Texas Instruments [25], Digi-Key [26], and Newark element14 [27], BeagleBoard is an open-source SoC computer [28]. It features TI Sitara AM5728 [29], which includes two C66x DSPs, two ARM Cortex-A15, two ARM M4 cores [30], and two PowerVR SGX544 GPUs [31]. With its relatively low price, the DSPs along with the co-processors make a powerful platform for implementing standalone SDRs.

4.3. General Purpose Processor-based:

1. Universal Software Radio Peripheral N-Series [32].
2. Ziria: A programming platform uses a domain specific language (DSL) Called Ziria.
3. Sora: A fully programmable software radio platform on PC architecture [33].
4. Lime Microsystems SDR: (Field Programmable RF transceiver (FPRF) technology).
5. Kansas University Agile Radio (KUAR): A Flexible Software-Defined Radio Development Platform. In addition, a Project for an M2M application platform that provides a Java/OSGi-based container for application running in service gateways. KURA covers I/O access, data service, watchdog service, network configuration and remote [64].

4.4. Graphics Processing Unit-based:

1. OFDM for Wi-Fi Uplink SDR.
2. WiMAX SDR.
3. Signal Detection SDR.

4.5. Hybrid Design (Co-design):

1. Wireless open-Access Research Platform (WARP) [36].
2. USRP Embedded (E) Series.
3. PSoC 5LP.
4. Zynq-based SDR.

4.6. SDR Reference guide

Table 2 compares a list of available software-defined radio indicated above and others more in an effort to provide a reference guide for developers according to some standards:

Table 2

Comparison between available SDR

SDR-Name	Frequency	Bandwidth	ADC (Bit)	DAC (Bit)	Sample rate	Interface	FPGA	Price US$
Apache Labs ANAN-8000DLE [37]	0 kHz - 61.44 MHz	x	16	16	x	Gigabit Ethernet	Altera Cyclone IV	4,395
AD-FMCOMMS5-EBZ [38]	70 MHz – 6 GHz	54 MHz	12	12	61.44 MSPS	FMC (to Xilinx board) then USB 2.0 or Gigabit Ethernet.		1,125
ADALM-PLUTO [39]	325 MHz – 3.8 GHz	20 MHz	12	12	61.44 MSPS	USB 2.0, Ethernet&WLAN with USB-OTG	Xilinx Zynq Z-7010	148
AFEDRI SDR [40]	30 kHz – 35 MHz, 35 MHz – 1700 MHz	2.3MHz	12		80 MSPS	USB 2.0, 10/100 Ethernet		249
Aaronia SPECTRAN V6 [41]	20 MHz – 6 GHz	Up to 490 MHz	16	16	2 GSPS	Embedded or True IQ data via 2 x USB 3.2 Gen1, 1 x USB 3.1 GEN2	XC7A200T-2 (930 GMACs)	3,400
AirSpy R2 [42]	24 – 1700 MHz	10 MHz	12	N/A	10 MSPS ADC sampling, up to 80 MSPS for custom applications	USB	None	170
AirspyHF+ [43]	9 kHz -31 MHz (60 -260) MHz	660 kHz	18	N/A	36 MSPS	USB		199
AOR AR-2300 [44]	40 kHz – 3.15 GHz		x	N/A	65 MSPS	Embedded system (no computer needed), USB		3,299
ASR-2300 [45]	300 MHz – 3.8 GHz		x	X	<40 MHz (Programmable)	USB 3.0 SuperSpeed		1,500
Bitshark Express RX [46]	300 MHz – 4 GHz		x		105 MSPS (RX only)	PCIe		4,300
bladeRF 2.0 micro [47]	47 MHz – 6 GHz	56 MHz	12	12	61.44 MSPS	USB 3.0 SuperSpeed	Altera Cyclone V	480
ColibriDDC [48]	10 kHz – 62.5 MHz	38 – 312 kHz	14	N/A	125 MSPS	10/100 Ethernet		650
COM-3011 [49	20 MHz – 3 GHz		Ext		External ADC required (I/Q output)	USB		345
Cyan [50]	100 kHz – 18 GHz	1 – 3 GHz	12 – 16	16	1–3 GSPS ADCs 2.5 GSPS DACs	4x 40Gbps QSFP, Ethernet	Intel Stratix 10 SoC	73,500
DRB 30 [51	30 kHz – 30 MHz		Ext		External ADC required (I/Q output)	LPT parallel port		390
DX Patrol [52]	100 kHz–2GHz		8		2.4 Msps	USB		115
ELAD FDM-S1 [53]	20 kHz–30 MHz	192- 3072 kHz	14	N/A	61.44 MHz	USB	Xilinx	420
ELAD FDM-S2 [54]	HF:9 kHz – 52 MHz / FM:74 MHz - 108 MHz / VHF:135 MHz - 160 MHz	192 kHz – 6 MHz	16	N/A	122.88 MHz	USB 2.0	Xilinx Spartan-6	600
ELAD FDM-DUO [55]	HF:10 kHz – 54 MHz	192 kHz – 6 MHz	16	X	122.88 MHz	Embedded system + 3x USB 2.0	Xilinx Spartan-6	1300
Elecraft KX3 [56]	0.5 – 54 MHz (144–148 MHz optional)		14	X	30 kHz	USB or embedded system (no computer needed)		900
FLEX-6700 [57]	0.01 – 73, 135 – 165 MHz	24-192kHz	16	16	245.76 MSPS	Gigabit Ethernet	Xilinx XC6VLX130T	7000
CDRX-3200 [58]	0.01 – 100 MHz	48 – 250 kHz	24	-	48-250 kSPS	Gigabit Ethernet	Xilinx XC5VLX30T	
LBRX-24 [59]	950 – 2150 MHz	150 kHz–80 MHz	16	-	150 KSPS – 80 MSPS	10 Gigabit Ethernet (x4)	Xilinx XC6VHX380T	
FLEX-6600M [60]	0.01 – 54 MHz	24 – 192 kHz	16	16	245.76 MSPS	Gigabit Ethernet	Xilinx XC6VLX130T or XC7A200T	5000
FLEX-1500 [61]	0.01 – 54 MHz	48 kHz	16	16	48 kHz	USB	-	650
Hermes-Lite2 [62]	0 to 38.4 MHz	1.536 MHz	12	12	76.8 MSPS	Ethernet	Altera Cyclone IV	300
Iris-030 [63]	50 MHz – 3.8 GHz	122.88 MHz	12	12	122.88 Msps (SISO) 61.44 Msps (MIMO)	Gigabit Ethernet or 24.6 Gbps High-Speed Bus	Xilinx Zynq 7030	2,400
KUAR Radio [64, 65]	5 GHz	TX 30 MHz RX 600 MHz.	6	16	160 Msps	Embedded PC built + ComExpress 1.4 GHz Pentium M	Xilinx Virtex II Pro P30	-
KerberosSDR [66]	24MHz - 1.7GHz	4* sample rate	8		2.4 Msps	USB		150
LimeSDR [67]	100 kHz – 3.8 GHz	61.44 MHz (120 MHz internally)	12	X	61.44 Msps	USB 3.0, PCIe	Altera Cyclone IV	800
LimeSDR-Mini [68]	10 MHz – 3.5 GHz	30.72 MHz	12	X	30.72 Msps	USB 3.0, PCIe	Altera MAX 10	160
LD-1B [69]	100 kHz – 30 MHz		Ext		External ADC required (I/Q output)	USB		285
Matchstiq [70]	300 MHz – 3.8 GHz		x	X	40 MSPS (RX/TX)	Embedded System or USB	Xilinx Spartan 6	4,500
MB1 [71]	10 kHz – 160 MHz	38 – 312 kHz	16	14	160 MSPS (RX), 640 MSPS (TX)	10/100 Ethernet, WLAN (optional)		5,600
Mercury [72]	0.1 – 55 MHz		x		122.88 MSPS	USB (via Ozy) or Ethernet (via Metis)		469
Myriad-RF [73]	300 MHz – 3.8 GHz		x		Programmable (16 selections); 0.75 – 14 MHz, Bypass mode	standard connector FX10A-80P	None	300
Noctar [74]	100 kHz – 4 GHz	200 MHz	x		x	PCI Express ×4		2,500
Odyssey TRX [75]	0.5 – 55 MHz		x		122.880 Msps ADC sampling, 48k-960k output sample rate	LAN, Wi-Fi, USB	Altera Cyclone IV	450
Perseus [76]	10 kHz – 40 MHz	1.6 MHz	16		80 MS/s (16 bit ADC)	USB 2.0		1200
PCIe SDR MIMO [77	70 MHz – 6 GHz		x		61.44 Msps	PCIe (1x)		1700

Table 2

Comparison between available SDR (prolongation)

SDR-Name	Frequency	Bandwidth	ADC (Bit)	DAC (Bit)	Sample rate	Interface	FPGA	Price US$
PM-SDR [78]	100 kHz – 50 MHz	192 kHz	Ext		External ADC	USB		220
PrecisionWave Embedded SDR [79	1 MHz – 9.7 GHz	2x RX: 155 MHz	x		310 MSPS	Embedded System Gigabit Ethernet / USB / JTAG / Audio	Xilinx Zynq Z-7030	4000
QS1R [80	10 kHz–62.5 MHz		x		130 MHz	USB	Altera Cyclone III	1010
Quadrus [81]	0.1 – 440 MHz		x		80 Msps ADC sampling, 48k-1.536M output sample rate	PCI		1550
Realtek RTL2832U DVB-T tuner [82]	24 – 1766 MHz (R820T tuner)	Matches sampling rate	8		2.8 MHz	USB		10
RDP-100 [83]	RX, 0 – 125 MHz; TX, 0–200 MHz		x		RX: 250 MSPS TX - 800 MSPS	Embedded System		x
RTL-SDR V3 Receiver Dongle [84]	0.5 – 1766 MHz	Matches sampling rate	8		2.4 MHz	2.4 MHz		21.95-25.5
SDRplay [85]	1kHz – 2 GHz	10 MHz	14		Two independent tuners, each with 11 built-in preselection filters. 3 antenna ports	USB	None	290
SDR-IQ [86]	0.1 kHz – 30 MHz		x		66.666 MHz	66.666 MHz		525
SDR-IP [87]	0.1 kHz – 34 MHz		x		80.0 MHz	Ethernet		3000
SDR-LAB SDR04 [88]	0.4 – 4 GHz		x		40 MHz	USB 3.0 SuperSpeed		x
SDRX01B [89]	50 kHz – 200 MHz		Ext		< 2 MHz External ADC required (I/Q output)	Ethernet or USB usually, but other interfaces are available in MLAB modular system		100
SDR Minor [90]	0.1 – 55 MHz		x		122.880 Msps ADC sampling, 48k-960k output sample rate	LAN 10/100		200
SDRstick UDPSDR-HF2 [91]	0.1 – 55 MHz		x		122.88 Msps	1G Ethernet via BeMicroCV-A9	Altera	400
SDR MK1.5 Andrus [92]	5 kHz – 31 MHz (1.7 GHz downconverter opt.)		x		64 MSPS	USB 2.0, 10/100 Ethernet		480
SDR-4+ [93]	0.85 – 70.5 MHz		x		48 kHz (integrated soundcard)	USB × 2		260
SDR(X) HF, VHF & UHF [94]	0.1 – 1850 MHz (R820T tuner)		x		Optimized for HF amateur bands with 4 user selectable pre-select HF filters	USB		100
SoftRock RX Ensemble II LF [95]	180 kHz – 3.0 MHz		Ext		External ADC required (I/Q output)	USB		97
Spectre [96]	0.4 – 4 GHz	200 MHz	16		310 MSPS	USB, Serial, JTAG, 10Gbit/s SFP+ Ethernet		10,000
SunSDR2 Pro [97]	10 kHz – 160 MHz	38 – 312 kHz	16	14	160 MSPS (RX), 640 MSPS (TX)	10/100 Ethernet, WLAN (embedded)		1,595
ThinkRF WSA5000 [98]	50 MHz – 8 GHz, 18 GHz or 27 GHz		x		125 MSPS	10/100/1000 Ethernet		3,500-14,140
USRP B210 [99]	70 MHz – 6 GHz	56 MHz	x		56 Msps	USB 3.0	Xilinx Spartan 6 XC6SLX150	1,100
USRP N210 [100]	DC – 6 GHz	Up to 40 MHz	16		25 Msps for 16-bit samples; 50 Msps for 8-bit samples	Gigabit Ethernet	Xilinx Spartan 3A-DSP 3400	1,717
USRP X310 [101]	DC – 6 GHz	Up to 160 MHz	x		200 Msps	Gigabit Ethernet, 10 Gigabit Ethernet, PCIe	Xilinx Kintex-7 XC7K410T	4,800
UmTRX [102]	300 MHz – 3.8 GHz	Up to 28 MHz	12	12	13 MSPS x2	Gigabit Ethernet	Spartan 6 LX75	1,300
WARPv3 [103]	2.4 GHz and 5.8 GHz	40 MHz	12	12	40 Msps	Dual Gigabit Ethernet	Xilinx Virtex-6 LX240T	7000
WinRadio [104]	9 kHz – 50 MHz		x	N/A	100 MSPS	USB		950
XTRX Pro [105]	30 – 3700 MHz	120 MHz	12	12	120 MSRP SISO, 90 MSRP MIMO	Mini PCIe	Xilinx Artix7 50T	599

5. Development Tools

In this paragraph, we offer the common development tools, which Researches typically use in the process of SDR design and implementation for different design approaches. We also provide an overall comparison between them to highlight the differences as shown in **Table 4.**

5.1. High Level Synthesis (HLS)

Table 3 presents a summary of HLS tools. The Tools in table all provide a set of area and timing optimizations such as resource sharing, scheduling, and pipelining. Nevertheless, not all of them are capable of generating testbenches for the design.

Table 3: HLS Tools

	Xilinx Vivado HLS	Intel FPGASDK OpenCL	Cadence Stratus HLS	Synopsys Synphony C Compiler	Maxeler MaxCompiler	MATLAB HDL Coder HLS
Input	C/C++/SystemC	C/C++/SystemC	C/C++/SystemC	C/C++	MaxJ	Algorithm & Modeling
Output	VHDL/Verilog/SystemC	VHDL/Verilog	VHDL/Verilog	VHDL/Verilog/SystemC	VHDL	VHDL/Verilog/SystemC
Test bench	Yes	No	Yes	Yes	No	Yes
Optimizations	Yes	Yes	Yes	Yes	Yes	Yes
Compatibility	Xilinx FPGA	Intel FPGA	All	All	All	All

5.2. Tools

1. MATLAB & HDL Coder.

2. LabVIEW.

3. GNU Radio.

4. Vivado HLS & SDSoC.

5. Compute Unified Device Architecture (CUDA) [106].

6. Joint Tactical Radio System (JTRS) [107]: Was a program of the US military to produce radios that provide flexible and interoperable communications. Examples of radio terminals that require support include hand-held, vehicular, airborne and dismounted radios, as well as base-stations (fixed and maritime). Allow radio components to be distributed across heterogeneous computer hardware, including FPGAs, DSPs, and GPPs. This goal is achieved using SDR systems based on an internationally endorsed open Software Communications Architecture (SCA). This standard uses CORBA on POSIX operating systems to coordinate various software modules. The program is providing a flexible new approach to meet diverse soldier communications needs through software programmable radio technology. All functionality and expandability is built upon the SCA.

Table 4

Development Tools & Platforms

	MATLAB & Simulink	Vivado HLS & SDSoC	LegUP	GNU Radio	Lab View	CUDA
Input	MATLAB/Graphical	C/ C++/ System C	C	Graphical/Python/C++	Graphical	C/C++/Fortran/Python
Output	MATLAB/C++/RTL	C/RTL	C/RTL	C/RTL	C/RTL	Machine Code
Platform	GPP/GPU/DSP/FPGA	GPP/FPGA	GPP/FPGA	GPP/GPU/DSP/FPGA	GPP/GPU/DSP/FPGA	GPU
License	commercial	commercial	open-source	open-source	commercial	commercial

The SCA, despite its military origin, is under evaluation by commercial radio vendors for applicability in their domains. The adoption of general-purpose SDR frameworks outside of military, intelligence, experimental and amateur uses, however, is inherently hampered by the fact that civilian users can more easily settle with a fixed architecture, optimized for a specific function, and as such more economical in mass market applications. Still, software-defined radio's inherent flexibility can yield substantial benefits in the longer run, once the fixed costs of implementing it have gone down enough to overtake the cost of iterated redesign of purpose built systems. This then explains the increasing commercial interest in the technology.

The Open Source SCA Implementation – Embedded (OSSIE [108] project, provides SCA-based infrastructure software and rapid development tools for SDR education and research. The Wireless Innovation Forum funded the SCA Reference Implementation project, an open source implementation of the SCA specification. (SCARI) can be downloaded for free.

According to what previously mentioned, we conclude that each SDR is unique concerning the design methodology, development tools, performance, and end application.

5.3. Case Study: (Hardware-Software Co-Design Workflow for System on Chip Platforms)

This paragraph Explain an academic Example Design supports our Article idea using MATLAB & Simulink & HDL Coder to develop an SDR with a desktop computer and SoC platforms. **Figure 3** shows the design flow for SoC platforms that the aforementioned tools offer and how they are connected.

The HDL Coder Hardware-Software Co-design workflow helps automate the deployment of MATLAB and Simulink design to a Zynq-7000

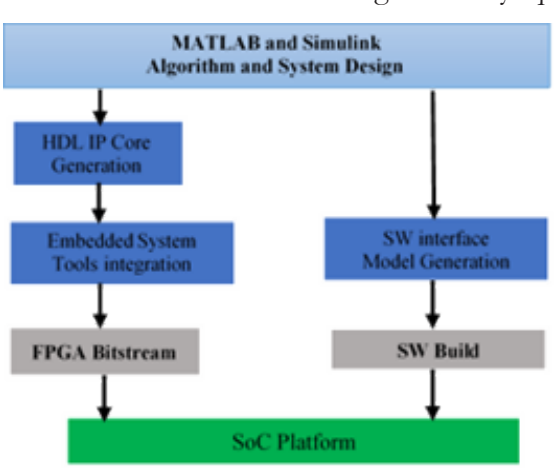


Fig. 3. *Hardware-Software Co-Design Workflow for SoC.*

platform or Intel SoC platform. We can explore the best ways to partition and deploy our design by iterating through the following workflow.

1. MATLAB and Simulink Algorithm and System Design: we begin by implementing our design in MATLAB or Simulink. When the design behavior meets requirements, decide how to partition our design: which parts we want to run in hardware, and which parts to run in embedded software. The part of the design that we want to run in hardware must use MATLAB syntax or Simulink blocks that are supported and configured for HDL code generation.

2. HDL IP Core Generation: Enclose the hardware part of our design in an atomic Subsystem block or MATLAB function, and use the HDL Workflow Advisor to define and generate an HDL IP core.

3. Embedded System Tool Integration: As part of the HDL Workflow Advisor IP core generation workflow, we insert our generated IP core into a reference design, and generate an FPGA bitstream for the SoC hardware.

The reference design is a predefined embedded system integration project. It contains all elements the Intel or Xilinx software needs to deploy our design to the SoC platform, except for the custom IP core and embedded software that we generate.

4. SW Interface Model Generation (requires a Simulink license and Embedded Coder license): In the HDL Workflow Advisor, after we generate the IP core and insert it into the reference design, we can optionally generate a software interface model. The software interface model is our original model with AXI driver blocks replacing the hardware part.

If the designer or researcher have an Embedded Coder license, he/she can automatically generate the software interface model, generate embedded code from it, and build and run the executable on the Linux kernel on the ARM processor. The generated embedded software includes AXI driver code generated from the AXI driver blocks that controls the HDL IP core.

However, if do not have an Embedded Coder license or Simulink license, he/she can write the embedded software and manually build it for the ARM processor. The following diagram shows the difference between the original model and the software interface model as shown in **Figure 4.**

5. SoC Platform and External Mode PIL: Using

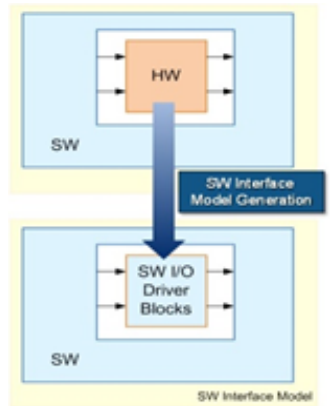


Fig. 4. *The difference between the original model and the software interface* [110].

the HDL Workflow Advisor, we program our FPGA bitstream to the SoC platform. We can then run the software interface model in external mode, or processor-in-the-loop (PIL) mode, to test our deployed design.

Finally, if our deployed design does not meet the design requirements, we should repeat the workflow with a modified model, or a different hardware-software partition.

6. FUTURE DIRECTIONS & CONCLUSIONS

We thing that future Directions of SDR Researches support the goals and key actions of the Europe 2020 initiative and the Digital Single Market and in particular focuses on:

- Eliminating the digital divide;
- Efficient use of spectrum;
- Promoting investments, competition and innovation; and
- Protecting general interest objectives such as cultural diversity and media pluralism.

In addition, on the EU level, radio spectrum policy has three main goals, which are:

- The harmonization of spectrum access conditions across the Union's internal market, enabling interoperability and economies of scale for wireless equipment.
- A more efficient use of spectrum;
- Improve availability of information about the current use, plans for use and availability of spectrum.

In our Research, we presented a comprehensive overview of the various design approaches and reconfigurable platforms adopted for SDR solutions. This includes GPPs, GPUs, DSPs, FPGAs, and Co-

design. We explained the basic architectures. Then reviewed the major current and early SDR platforms, whether they were developed by the industry or in academia; Due to the different features of design approaches and development tools, we found that it was important to compare them against each other based on a set of metrics as a guide to developers. Finally, we offered some of the SDR research challenges and topics that are predicted to improve in the near future, helping to advance SDRs and their wide adoption.

Finally, we think that SDR solutions are going to be mainstream and that their ability to implement different wireless communication standards with high levels of flexibility and reprogrammability will be considered the norm. In a few last words this Research concluded to: that the driving factors for the high demand of SDR include network interoperability, readiness to adapt to future updates and new protocols, and more importantly, lower hardware and development costs.

REFERENCES

1. Cisco–Global Home Page, www.cisco.com, 2020.
2. Buyer. Software Defined Radio Market by Application, Component, End User, Type - Global Forecast to 2021. Tech. Rep., 2016. [Online]. Available: https://www.reportbuyer.com.
3. A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Communications Surveys and Tutorials*, 2015.
4. Global Industry Analysts Inc. Software Defined Radio (SDR) - Global Strategic Business Report. Tech. Rep., 2017. [Online]. Available: https://www.researchandmarkets.com/research/pc9x7g/software defined.
5. Dale J. Mortensen, Daniel W. Bishop. Space Software Defined Radio Characterization to Enable Reuse. *JAN*, 2020.
6. Gara Quintana-Diaz, Roger Birkeland. *Software Defined Radioi IN SATELLITE COMMUNICATIONS*, Research-Gate, Jan 2019.
7. J. Seo, Y.-H. Chen, D.S. De Lorenzo, S. Lo, P. Enge, D. Akos, and J. Lee. A real-time capable software-defined receiver using GPU for adaptive anti-jam GPS sensors. *Sensors,* 2011, 11(9):8966-91.
8. W. Xiang, F. Sotiropoulos, and S. Liu. *Radio: An Novel Software Defined Radio (SDR) Platform and Its Exemplar Application to Vehicle-to-Vehicle Communications*. Springer, Cham, 2015, pp. 404-415.
9. M. Kloc, R. Weigel, and A. Koelpin. SDR implementation of an adaptive low-latency IEEE 802.11p transmitter system for real-time wireless applications. *IEEE Radio and Wireless Symposium,* Jan 2017, pp. 207-210.
10. Y. Chen, S. Lu, H.-S. Kim, D. Blaauw, R. G. Dreslinski, and T. Mudge. A low power software-defined-radio baseband processor for the Internet of Things. *IEEE International Symposium on High Performance Computer Architecture* (HPCA-2016), mar 2016, pp. 40-51.
11. Y. Park, S. Kuk, I. Kang, and H. Kim. Overcoming IoT Language Barriers Using Smartphone SDRs. *IEEE Transactions on Mobile Computing*, 2017, 16(3):816-828.
12. https://Imagine.gsfc.NASA.gov.
13. *IEEE Project 1900.1* - Standard Definitions and Concepts for Dynamic Spectrum Access: Terminology Relating to Emerging Wireless Networks, System Functionality, and Spectrum Management. https://standards.ieee.org/develop/project/1900.1.html.
14. A. Haghighat. A review on essentials and technical challenges of software defined radio. *Proceedings MILCOM*, 2002, pp. 377-382.
15. U.L. Rohde and T.T.N. Bucher. *Communications Receivers: Principles and Design*. McGraw-Hill Education, 1989.
16. T.J. Rouphael. *RF and digital signal processing for software-defined radio: a multi-standard multi-mode approach*. Newness, 2009.
17. J.J. Carr. *The technician's radio receiver handbook: wireless and telecommunication technology*. Newness, 2001.
18. R. Walden. Analog-to-digital converter survey and analysis. *IEEE Journal on Selected Areas in Communications*, 2000, 17(4):539-550.
19. T. Hentschel, M. Henker, and G. Fettweis. The digital front-end of software radio terminals. *IEEE Personal Communications*, 2000, 6(4):40-46.
20. C. Bowick, J. Blyler, and C. J. Ajluni. *RF circuit*

design. Newness/Elsevier, 2012.
21. M.N.O. Sadiku and C.M. Akujuobi. Software-defined radio: a brief overview. *IEEE Potentials,* 2004, 23(4):14–15.
22. M.M. Bezem, J.W. Klop, R.de Vrijer. *Term rewriting systems*. Cambridge University Press, 2003.
23. B. Khailany, W.J. Dally, U.J. Kapasi, P. Mattson, J. Namkoong, J.D. Owens, B. Towles, A. Chang, and S. Rixner. Imagine: media processing with streams. *IEEE Micro*, 2002, 21(2):35-46.
24. B.K. Khailany, T. Williams, J. Lin, E.P. Long, M. Rygh, D.W. Tovey, and W.J. Dally. A Programmable 512 GOPS Stream Processor for Signal, Image, and Video Processing. *IEEE Journal of Solid-State Circuits*, 2008, 43(1):202-213.
25. *SMJ320C80 Digital Signal Processor - TI.com*. [Online]. Available: http://www.ti.com/product/SMJ320C80.
26. Digi-Key Electronics - Electronic Components Distributor. [Online]. Available: https://www.digikey.com.
27. *Newark element14 Electronics – Electronic Components Distributor*. [Online]. Available: https://www.newark.com
28. A.S. Fayez. Designing a Software Defined Radio to Run on a Heterogeneous Processor. P*h.D. dissertation, Virginia Tech*, Apr. 2011.
29. D.A. Patterson and J.L. Hennessy. *Computer organization and design ARM Edition: The Hardware Software Interface*. Morgan Kaufmann Publ., 2016.
30. *Architecting a Smarter World Arm*. [Online]. Available: https: //www.arm.com.
31. I*magination Technologies - Developing and Licensing IP cores*. [Online]. Available: https://www.imgtec.com.
32. *Ettus Research-Networked Software Defined Radio (SDR*). [Online]. Available: https://www.ettus.com.
33. K. Tan, H. Liu, J. Zhang, Y. Zhang, J. Fang, and G.M. Voelker. Sora: high-performance software radio using general-purpose multicore processors. Communications of the ACM, 2011, 54(1):99.
34. M. Bansal, A. Schulman, and S. Katti. Atomix: A Framework for Deploying Signal Processing Applications on Wireless Infrastructure. *12th USENIX Symposium on Networked Systems Design and Implementation* (NSDI 15), 2015, pp. 173–188.
35. M.-C. Ng, K. Fleming, M. Vutukuru, S. Gross, Arvind, and H. Balakrishnan. Airblue: A system for cross-layer wireless protocol development. *Symp. Architectures for Networking & Communications Syst.* (ANCS), pp. 1–11, 2010.
36. *WARP Project*. [Online]. Available: https://warpproject.org.
37. *Apache Labs*. Www. apache-labs.com. 2018.
38. http://www.analog.com/ad-fmcomms5-ebz http://wiki.analog.com/resources/eval/user-guides/ad-fmcomms5-ebz http://www.analog.com/ad9361.
39. http://www.analog.com/media/en/news-marketing-collateral/product-highlight/ADALM-PLUTO-Product-Highlight.pdf.
40. *Archived copy*. Archived from the original on 2013.
41. *SPECTRAN V6*. Www.aaronia.com. January 2020.
42. *Airspy SDR#*|Low Cost High Performance Software Defined Radio. Www.Airspy.com.
43. *Airspy HF+*. Www.Airspy.com. Retrieved 2018.
44. AR2300|RECEIVERS|AOR U.S.A., INC. *Authority On Radio Communications*. Www.Aorusa.com.
45. *Archived copy*. Archived from the original on 2014.
46. *Bitshark Express RX|Epiq Solutions*. Www.Epiqsolutions.com. July 2016.
47. *Nuand|bladeRF 2.0 micro*. Www.Nuand.com. Nov 2018.
48. *Expert Electronics - ColibriDDC*. Www.Eesdr.com. Dec 2017.
49. *COM-3011 [20 MHz - 3 GHz] Receiver/SDR*. Www.Comblock.com. 2020.
50. *Per Vices Home – Per Vices*. Www.Pervices.com. Feb 2019.
51. *Software Defined Radio - NTi Rudolf Ille Communications Technology - Products - DiRaBox*. Online 2020.
52. Www.dxpatrol.pt. 2020.
53. *FDM-S1 Receiver*. Www.Ecom.eladit.com. 2020.
54. *ELAD FDM-S2 SDR Receiver*. Www.Ecom.eladit.com. 2020.
55. *FDM-DUO SDR TRANSCEIVER*.www.Eecom.eladit.com.
56. *Elecraft® Hands-On Ham Radio™*. www.Elecraft.com.
57. *FLEX-6700-FlexRadio Systems*. Www. flexradio.com.
58. *CDRX-3200-FlexRadio Systems*. Www. flexradio.com. 2019.

59. *LBRX-24-FlexRadio Systems*. Www. flexradio. com.
60. *Lunaris SDR based on HERMES SDR Transceiver design*. www.Ceda-labz.com.
61. *FLEX-1500-FlexRadio Systems* .Www. flexradio. com.
62. *HL2 build9 specs*. 2019.
63. *Products-Skylark Wireless*. 2019
64. https://www.eclipse.org/proposals/tecnology. Kura. Sep 2016.
65. G.J. Minden, J.B. Evans, L. Searl, D. DePardo, V.R. Petty, R. Rajbanshi, T. Newman, Q. Chen, F. Weidling, J. Guffey, D. Datla, B. Barker, M. Peck, B. Cordill, A.M. Wyglinski and A. Agah. *KUAR: A Flexible Software-Defined Radio Development Platform*. The University of Kansas, Lawrence, KS 66045.
66. *KerberosSDR - 4 Channel Coherent RTL-SDR*.
67. *LimeSDR*: Flexible, Next-generation, Open Source Software Defined Radio, Crowd Supply. Www.limesdr.org.
68. *LimeSDR-Mini*: Flexible, Next-generation, Open Source Software Defined Radio, Crowd Supply.
69. *LD-1B Software-Defined Radio*.
70. *Matchstiq s10*. Www. Epiqsolutions.com.
71. *Expert Electronics - MB1*. Www.Eesdr.com.
72. *TAPR Webmaster*, Site Design by Greg Jones, WD5IVD. "TAPR - HPSDR Mercury". Www. Tapr.org.
73. *Reference Development Kit-Myriad*. Www.Myriadrf. org.
74. *Archived copy*.
75. *Odyssey,* New open source 16-bit HF DDC SDR Transceiver Odyssey|New open source 16-bit HF DDC SDR Transceiver. Www.Ody-sdr.com.
76. *Perseus SDR Home Page*. Www.Microtelecom.it.
77. *Archived copy* (PDF).
78. *IW3AUT-HAM RADIO PROJECTS*. Www. Iw3aut.altervista.org.
79. *PrecisionWave SDR*. www.precisionwave.com.
80. Software Radio Laboratory LLC. *QS1R Software Defined Receiver*. Www.Srl-llc.com.
81. QUADRUS SDR hardware digitizer|*SDR software receiver*. Www.Spectrafold.com.
82. http://sdr.osmocom.org/trac/wiki/rtl-sdr http://www.rtlsdr.com/ http://www.rtl-sdr. com.
83. *Archived copy*.
84. *Buy RTL-SDR Dongles* (RTL2832U). Www.Rtl-sdr.com.
85. *SDRplay*. www.sdrplay.com.
86. *SDR-IQ Receiver*. Www. Rfspace.com.
87. *SDR-IP*. Www.Rfspace.com.
88. http://www.sdr-lab.com http://amitec.co.
89. http://wiki.mlab.cz/doku.php?id=en:sdrx http://www.ust.cz/shop/product_info. php?cPath=29&products_id=76.
90. КВ приемник SDR-Minor - Мои статьи - *Каталог статей - Персональный сайт*. Www.Sdr-deluxe.com.
91. *SDRstick*. Www.Sdrstick.com.
92. *UVB-76 Live Stream Blog*. Uvb-76.net.
93. *Cross Country Wireless SDR-4+ general coverage receiver*. Www.Crosscountrywireless.net.
94. *Welcome to 6V6 Electronics - 6V6 Electronics Company*.
95. *SoftRock RX Ensemble II LF Receiver Kit*. Www. Fivedash.com.
96. *Spectre|Clearbox Systems*. Www.Clearboxsystems. com.au.
97. *Expert Electronics - SunSDR2 Pro*. eesdr.com.
98. *WSA5000|ThinkRF*. Www.Thinkrf.com.
99. USRP B210 USB Software Defined Radio (SDR) - *Ettus Research*. Www.Ettus.com.
100. USRP N210 Software Defined Radio (SDR) - *Ettus Research*. Www.Ettus.com.
101. USRP X310 High Performance Software Defined Radio (SDR) - *Ettus Research*. Www. Ettus.com.
102. *Google Code Archive* - Long-term storage for Google Code Project Hosting.
103. *Mango Communications - WARP v3 Kit. Www*. Mangocomm.com.
104. *WiNRADiO WR-G31DDC 'EXCALIBUR' Receiver*. Www.Winradio.com.
105. XTRX - A Fairwaves tiny SDR. *XTRX - A Fairwaves tiny SDR*.
106. CUDA Toolkit Documentation. [Online]. Available: *http://docs. Nvidia.com/CUDA*.
107. Dsca.mil/major-arms-scales/Canada-multifunctional-information-distrbution-System-*JTRS*.
108. https://www.openhup.net.
109. https://ieeexplore.ieee.org.
110. www.MathWorks.com.

DOI: 10.17725/rensit.2020.12.219

Chaotic signal processing and generation in DRFM technologies: accounting for resource constraints

Yuri N. Gorbunov

Kotelnikov Institure of Radioengineering and Electronics of RAS, Fryazino branch, http://fire.relarn.ru/
Fryazino 141190, Moscow region, Russian Federation
A.I. Berg Central Research Institute of Radioengineering, http://www.cnirti.ru/
Moscow 107078, Russian Federation
E-mail: gorbunov@ms.ire.rssi.ru

Gurgen L. Akopyan

A.I. Berg Central Research Institute of Radioengineering, http://www.cnirti.ru/
Moscow 107078, Russian Federation
E-mail: akopyan@cnirti.ru

Received March 12, 2020, reviewed March 30, 2020, accepted April 10, 2020

Abstract. **DRFM technology for storing radio frequencies does not require a large bit and is compatible (regular way) with the requirement to account for resource constraints (hardware and computing), but this applies to a single-signal situation. Significant complicates of signal processing arise in multi-signal situations at a large dynamic range: parasitic combination components appear, efficient separation (resolution) of signals is difficult. The article establishes that there is an alternative DRFM under construction device in such conditions, and it is digital multichannel filtering (DMF) of signals implemented in the device itself. However, if multiple bit processing of the representation and the current digital data is maintained, the multi-signal processing is greatly complicated. In order to reduce the effect of quantization and signal sampling effects, the article proposes to apply an unconventional approach, which is based on chaotic processing - randomization of tough ("low-bit") signal samples in the ADC. In addition, it has been found that in order to reduce DRFM discharge requirements, it is advantageous to apply a procedure for digitally subtracting the dominant signal from the input mixture, which accounts for a significant portion of the range (discharge).**

Keywords: **digital radiofrequency memory DRFM, digital multi-channel filtering, randomization, multi-signal mode, digital cut-off filtering, stochastic linearization, low-bit processing, coarse statistics, dominant signal**

UDC 621.396.96

For citation: Yuri N. Gorbunov, Gurgen L. Akopyan. Chaotic signal processing and generation in DRFM technologies: accounting for resource constraints. *RENSIT*, 2020, 12(2):219-226; DOI: 10.17725/rensit.2020.12.2.219.

Contents

1. INTRODUCTION AND PROBLEM DEFINITION

Well known that DRFM technology was developed in 80s and found broad application in the systems of a radar-location, radio navigation, radio-electronic counteraction and a radio communication [1-9]. The priority belongs to works [1,6,8,9]. The DRFM technology allows to carry out storing and repeated reproduction of radio signals in a strip of frequencies several

hundreds of megahertz, with dlitelnost of signals tens of milliseconds and in the big dynamic range.

At the same time experience of use of DRFM technology revealed a number of problems which solution is defined by the prospects of its development and emergence of new technologies. The main problem consisted in obtaining digital copies of signals in the wide frequency and dynamic ranges.

Usually, the bearing frequency of entrance signals a priori is unknown, only frequency range, for example in the eighties, it is ~4÷18 GHz is often known. From the moment of emergence of DRFM for the last 40 years, in connection with continuous expansion of range of distribution of frequencies of the suppressed DEN, the main efforts of developers were directed to increase in operating range ΔF_p of frequencies of DRFM, i.e. the problem was constantly aggravated.

It is known that the operating range of DRFM is defined by the frequency of sampling of signals in time and this dependence is set by expression:

$$\Delta F_p \leq \Delta F_d,$$

where $F_d = 1/T$ – frequency, and T – an interval (period) of sampling of a signal in time. Consequently, the only way of expansion of operating range of frequencies of the DRFM device is increase in its frequency of sampling which is 600 MHz today, and the GHz is required 600÷3000. Range of the remembered signals with $\Delta F_p \leq \Delta F_d$ = 3000 MHz is sufficient F for the majority of technical applications.

Power of entrance signals is also a priori unknown and can change in the range up to 60÷70 dB In these conditions there is a problem of transformation of signals to numeric words, their subsequent storing and investment of the created signals with interfering modulation, their transformation from "figure" to an analog form and strengthening.

It is known [10], that with the fixed signal amplitude for achievement of level of parasitic components of a signal no more minus 30 dB it is necessary to transform to two not less than 4 categories of (bit) long everyone. Total length of the numeric word has to be not less than 8 bits, and the volume of the memory device of the DRFM device has to be $8 \cdot \Delta F_p \cdot W$ and bit where W – radio signal duration.

If amplitude of an entrance signal changes in the range up to 60 dB (10^3 time), then for achievement of an identical step (increment) of quantization Δ quadrature with any amplitude of a signal it is necessary to increase word length up to 18÷20 of categories, and STORAGE volume to (18÷20) $\Delta F_p \cdot \tau_i$ and bit. Compression of dynamic range by means of the amplifier with logarithmic characteristic allows to squeeze dynamic range in ~100 times and to respectively reduce length of the numeric word to 12 bits. The serious constraint reduces the length of the digital words to 8÷10 bits. In both cases, additional parasitic components of the spectrum of the reproduced signal on the harmonics of the carrier frequency are generated. Serious difficulties arise in the case of wide carrier change bands of ~100÷500 MHz or more when several signals at different frequencies enter the DRFM operating band. In this case, signal restriction generates additional parasitic signal components, which cannot be eliminated without special measures.

Method of dynamic range compression with the help of AGC circuits does not eliminate problems of multi-signal situation, as well as method of separate amplitude and phase conversion [10]. Also, traditional approach the range width which it is commensurable with a working strip of frequencies ΔF_p and it does not solve a problem of storing and reproduction of broadband signals. Signal restriction produces additional parasitic components in the spectrum of the reproduced signal, the levels of which may be significant. This also applies to coarse

quantization, which in the first approximation prevents such a transition.

It is generally accepted that a large dynamic range of processed signals and A/D converter bit are interconnected. On noise of the receiver σ it is taken away 1-3 category ADC, consider quantization commotion SD $\sigma_\Delta = \Delta / 2\sqrt{3} \approx 0.6\Delta$, where Δ - the price of the younger category ADC and at big word length of the L of ADC (L >> 1) neglect commotion of quantization. For F_d = 3 GHz at L = 12 we have the forecast of volume of the memory device at quadrature processing $2 \cdot 12 \cdot \Delta F_p \cdot \tau_i$ and bit on an entrance at 3 selections of two 12-bit words approximately for 1 nanosecond.

Due to the widespread introduction of the TO methods, emergence PLD and DSP, an opportunity to redistribute strengthening of a path, to carry out the digital multichannel filtration (DMF), to make narrower strips of certain canals and, thus - to increase extreme sensitivity opened (even for short impulses from 10÷50 of nanosecond).

Regardless of the counting system, the rounding method, the representation of digital data in "integers," with a "fixed" or "floating" comma, simply increasing the sampling rate of the ADC F_d to 3 GHz results in a substantial increase in the bit rate resulting in "bit redundancy" in the TO path.

2. DESCRIPTION OF "BIT REDUNDANCY" EFFECT

Fig. 1 shows the dependence of the increase in the bit capacity of the ADC on the duration of the stored radio signal W from 0 to 1.5 μs (accordingly increase of number of samples from 1 to 2^{13} and more) At F_d ≈ of 3 GHz, which results (in limit) in output bit increase to 21 (the lower straight line) and possibilities of realization of dynamic range of a path up to 120 dB that on 70÷100 dB exceeds necessary, i.e. is superfluous.

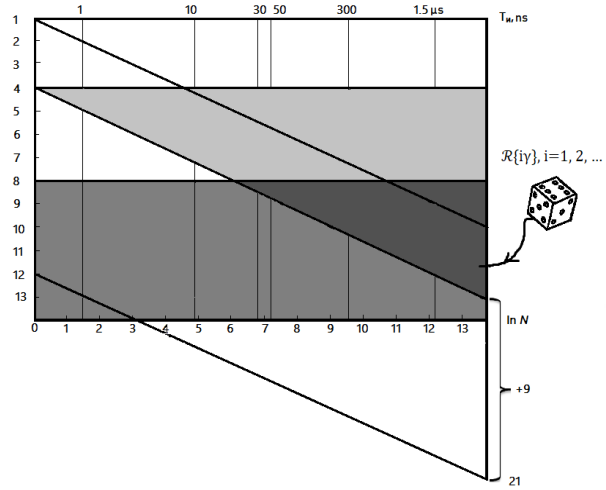


Fig. 1. *Diagram of increase of ADC converter digit capacity L depending on pulse duration W.*

On the top straight-line growth (on the module) is shown to word length for binary ("it is binary – sign") quantization below – 4-bit, allowing to have the output dynamic range of 50÷80 dB sufficient for high-quality reproduction of the digital copy of a signal. Simply increasing the bit rate during processing (without taking special measures) does not lead to reduction of quantization noise, because in a system with "fixed point" at accumulation of quantization error (in the absence of receiver noise) have (from count to count) the same value and sign (+ or –) and at summation in the TF ("accumulate," i.e. are not mutually compensated), and there is no improvement in the quantization signal-to-noise ratio.

In order to maintain the required dynamic range of the CP path under "rough quantization" conditions, it is proposed to select commotion from the following principle condition (1) (described by the expression):

$$\sigma << \Delta. \quad (1)$$

This means, that current intermediate counts "can be" rough. Weak signals, commensurate or smaller quanta, in the first approximation can be lost (inside the quant), and in other situations, on the contrary – to be emphasized (at the boundaries of the quanta).

3. CHAOTIC PROCESSING WITH LOW DIGIT CAPACITY: ROUGH STATISTICS

The chaotic DCP has been applied before. Consequently, stochastic ADCs and other structures were considered in [11]: it did not address special issues related to reducing sampling and quantization noise in DRFM devices.

In work [11] it is proposed to use stochastic interpolation of "rough" range samples to measure the range. When measuring target coordinates in this manner (essentially by the Monte-Carlo method), an interpolation task such as range D has been reduced to an interpolation additive measurement task Δ_x by measuring the associated probability $p = \Delta_x/\Delta$, where Δ is an element of discreteness, a quantization step by the measured parameter x.

In step [12-13], analysis of processing and generating signals in the radars with VTs was performed using randomization techniques. Solutions to eliminate uncertainty in the setting of probe signal parameters and in the selection of methods of processing and decision-making under resource constraints in VTs systems are justified, but this issue is solved in order to miss the suppression of passive interference in IR radio interference.

The works [14-19] consider a stochastic approach to solving traditional radar problems: detection, evaluation, filtering. Stochastic radar [14] is based on the concept of introducing artificial stochasticity into the process of processing and generation of radar signals, which assume, along with natural stochasticity due to random nature of input signals, randomization of reception-transmission process conditions. The solutions given therein are analogs.

The theoretical base of stochastic radar, and similar to the proposed approach, is the Monte-Carlo method. Further we will understand an essence of this method in relation to a solvable task in which signals in delta Δ are not excluded from processing, and we turn into some probability. Analogy of problem solution (by VT_S systems type in radar [13]) also determined the necessity of rational use of dynamic range, with respect to analyzed signal, and - subtraction of interference (dominance) before processing.

4. COMPENSATION OF THE DOMINANT SIGNAL BY "IF-CP " TYPE OF CIRCUIT IN STS SYSTEM

In the case under consideration, the injector filter (IF) is designed to compensate (subtract) the dominant signal (passive interference) from the sum of signals in a multi-signal situation in order to reduce the dynamic range of the MMF circuit made in [13] in the form of a multi-channel coherent storage device (DRFM of the total signal) with their subsequent separation.

The object of the CO in this pattern is the vibration coming from the receiver output $\dot{x}$, which is the sum of the analyzed signal $\dot{S}$ and the dominant signal $\dot{C}$.

The input signals for the ADC are the quadrature components x_c and x_s - respectively, the real (cosine) and imaginary (sine) parts of the complex vector $\dot{x}$.

A distinctive feature of the RO(PO) circuit is the presence of a random additive generator (RAG) designed to generate a random noise voltage $\vec{\xi}_{r+N}$ additive with elements $\vec{\xi}_i$, i = 1,2,...,r+N. The coherent store KN-TMF, distributes a set of harmonicas of "a group signal" on the separate "streamlets" forming "narrow" channels in a working strip ΔF_p can be realized on DPF algorithm:

$$z = \sqrt{f_c^2 + f_s^2}, \tag{2}$$

$$\dot{f}(k) = f_c + jf_s = \sum_{i=0}^{N-1} \dot{x}_i e^{-jia_k}, \tag{3}$$

where $a_k = 2\pi k/N$ - setting of k-th channel for inter-period run of signal phase from target; N - number of analyzed pulses in the packet and simultaneously (for DMF) number of frequency channels; k = 0,1,2,...,N-1 - channel number;

$\dot{x}_{iP\Phi} = x_{ci} + jx_{si}$ - temporary quadrature samples of signal at RF output.

As an indicator of system efficiency, we use the improvement factor [15]:

$$J = r_{out}/r_{in} = K_G/K_C, \quad (4)$$

where $r_{out} = P_{Cout}/P_{Gout}$, $r_{in} = P_{Cin}/P_{Gin}$ is the ratio of the power of the useful signal to the power of the dominant signal at the output and input, respectively; $K_G = P_{Gin}/P_{Gout}$ is the suppression factor of the dominant signal; $K_C = P_{Cin}/P_{Cout}$ is the transmission factor of the analyzed signal.

First, we will analyze the passage of the dominant signal. **Fig. 2** shows the canonical diagram of one quadrature channel of RF of r order, which for binomial weighting coefficients

$$a_i = (-1)^i C_r^i (i = 0,1,2,...,r), \quad (5)$$

where C_r^i is the number of combinations from r to i, identical to the scheme *r-FIR*.

In deterministic quantization, the current digital count is related to the level of compensated interference (dominant signal) ratio (Fig. 2, $\vec{\xi} = 0$):

$$C = X\Delta + \Delta_C,$$

where $X = E\{C/\Delta\}$ is the function of the whole part, $\Delta_C = R\{C/\Delta\}$ is the fractional fraction of the relation C/Δ.

Let $\vec{\xi} \in [0, \Delta]$, then at the output of ADC digital samples are generated $X + \mu_i$, i = 1,2..., where

$$\mu_i = \begin{cases} 1, \text{ with probability } p = \Delta_C/\Delta, \text{ for } \xi > \Delta - \Delta_C; \\ 0, \text{ with probability } q = 1 - p, \text{ for } \xi \le \Delta - \Delta_C. \end{cases}$$

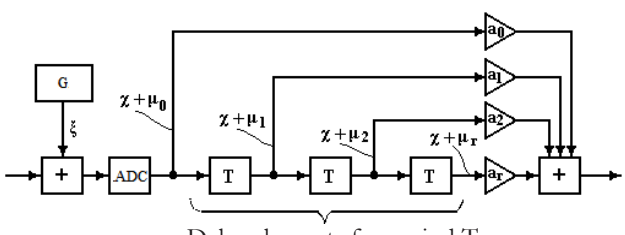


Fig. 2. *Canonical scheme of stochastic IF r-th order.*

Using the accepted symbols, the power of the dominant signal P_G at the output of the RF is represented as:

$$P_s = M\left\{\left[\Delta\sum_{i-0}^{r}(-1)^i C_r^i\left(\frac{C-\Delta_C}{\Delta}+\mu_i\right)\right]^2\right\}, \quad (6)$$

where M{...} is the mathematical expectation operator.

At independent tests $M\{\mu_i\} = p$, $M\{\mu_i^2\} = p^2$, $M\{\mu_i\mu_j\} = M\{\mu_i\}M\{\mu_j\} = p^2$, get

$$P_G = \Delta^2 pq\sum_{i=0}^{r}(C_r^i)^2 = P_{GO}. \quad (7)$$

Signals from the output of the dominant signal compensator in the DMF unit are subjected to transformations (1) and (2). Considering (for simplicity) the operation of one ("central") channel numbered $k = N/2$ (N is even), the expression for power deterministic at the output of the N-point DFT block:

$$P_{Gout} = M\left\{\left[\Delta\sum_{j=1}^{N}\left(\sum_{i=1}^{r}(-1)C_r^i\left(\frac{C-\Delta_C}{\Delta}+\mu_{j-i}\right)\right)\right]^2\right\}, \quad (8)$$

Because for this channel $\alpha_k = \pi$, $\sin j\pi = 0$, $\cos j\pi = (-1)^i$.

Without disturbing the commonality for the same frequency channel, the expression (8) taking into account (7) is converted to a view

$$P_{Gout} = \Delta^2 Npq\sum_{i=0}^{r}(C_r^i)^2 = NP_{GO}.$$

Considering further that the maximum value of the dominant signal is $C = \Delta 2^{i-1}$, and also that the amplitude of the analyzed signal, for example at the Nyquist frequency ($d_k = \pi$)v, after passing through the IF and the CH is increased by a factor $2^r N$, it is desirable to characterize the degree of suppression of the dominant signal for the PO processing by a normalized suppression factor.

$$K_{GNR} = \frac{\Delta^2 2^{2(L-1)}(2^r N)^2}{P_{GO}N} = \frac{2^{2(L-1)}2^{2r}N}{pq\sum_{i=0}^{r}(C_r^i)^2}. \quad (9)$$

The minimum value of the suppression factor is achieved for the interference lying in the middle of the quant (for them) ("antinode" are formed):

$$K_{GNRM} = 2^{2L} \frac{2^{2r} N}{\sum_{i=0}^{r} (C_r^i)^2} = 2^{2L} \eta. \quad (10)$$

At the determined quantization the dominating signals lying in quantum completely are suppressed, and the dominating signals lying on its borders are suppressed to a lesser extent as the level of not compensated remains at the exit of the IF and KN can reach the size $N2^{(r-1)}\Delta$. Therefore, the value of the normalized suppression factor for deterministic processing:

$$K_{GNDM} = \frac{\Delta^2 2^{2(L-1)} (2^r N)^2}{N^2 2^{2(r-1)} \Delta^2} = 2^{2L}. \quad (11)$$

Next, in expression (8), the coefficient $\eta > 1$, i.e., PO, has advantages over deterministic processing (DP). Actually $\left(\sum_{i=0}^{r} C_r^i \right)^2 > \sum_{i=0}^{r} (C_r^i)^2$, since, a $\sum_{i=0}^{r} C_r^i = 2^r$, we get:

$$\eta = \frac{2^{2r} N}{\sum_{i=0}^{r} (C_r^i)^2} > \frac{2^{2r}}{\left(\sum_{i=0}^{r} C_r^i \right)^2} = \frac{2^{2r} N}{2^{2r}} = N \geq 1. \quad (12)$$

Analysis of expression (9) shows that degree of suppression of dominant signals in case of RO is determined not only by bit L of ADC, but also by order r of IF, as well as by number **N** of accumulated samples in DMF unit. By selecting N and r respectively, the number of quantization levels of the input ADC can be significantly reduced $M = 2^L$ to achieve the desired suppression. Programming on the PLD divides the CP into two units: IF and CP the implementation of which is significantly simplified due to the sharp reduction of the CP discharge. In DO processing, as seen in (19), the degree of suppression is determined by ACD converter discharge L, wherein the specific suppression per bit does not exceed 6 dB.

5. ANALYSIS OF SIGNAL TRANSMISSION

The passage of weak signals commensurate with quantum and less was analyzed. The nonlinearity of the step amplitude characteristic can show results in that, if the amplitude of the useful signal $S < Q\{x/\Delta\}\Delta$, where $x = C \pm S$, $Q\{x/\Delta\}$ is a function of the distance to the nearest integer x/Δ, such a signal is lost during processing due to nonlinearity of the "dead zone" type. Randomization of the processing allows linearizing said nonlinearity and thus detecting the signal from the target located inside the quant of the ADC [13].

The power of the useful signal at the output of the ADC+N-Surgical DMF processing device was defined as the increment of the sum power, which is caused by the analyzed signal. It is shown that at PO the power of the analyzed signal, even if it is inside the quantum Δ ($B = 0$), at the output of the DC device is not equal to zero. Absence of "dead zone" in amplitude characteristic of RP device is explained by effect of "linearization of nonlinearity" of ADC.

Linearized characteristic of ADC is shown in **Fig. 3**: Dependencies are built on the amplitude P_{Cout} for the analyzed signal S at $\Delta_C = 0$ (curve 1) and $\Delta_C = \Delta/2$ (curve 3). The same figure shows that the corresponding constraints at NO (curves 2 and 4). As can be

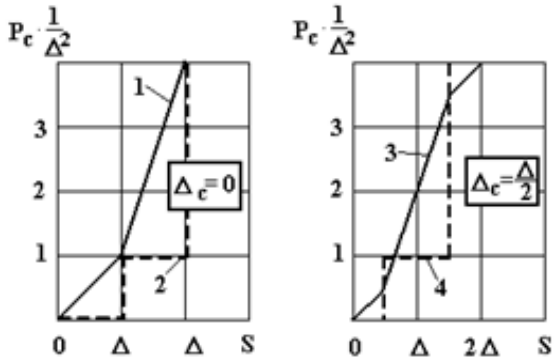


Fig. 3. *Relations of normalized power of output signal S, commensurate with quant size for compared CA methods.*

seen from the figure (curves 1 and 3), the effect of eliminating the "dead zone" in which weak signals were lost, was found. Consequently, linearization of non-linearity of the "dead zone" type allows to detect weak signals, the amplitude of which is commensurate and less than quantum ⊿.

If the amplitude of the useful signal is not too small compared to ⊿, its power at the output of the RO device:

$$P_{Cout} = \frac{(N2^r S)^2}{2} = N^2 2^{2r-1} S^2.$$

Given that $P_{Cin} = S^2/2$, the gain of the useful signal is represented by

$$K_C = P_{Cout} / P_{Cin} = N^2 2^{2r}. \quad (13)$$

It follows from the obtained formulas that the specified coefficient of improvement of the SDT-P filter with the corresponding selection of parameters *N* and *r*, IF and CD can be achieved with less than the deterministic processing of the number of quantization levels in the input ADC. Gain (equivalent) digit capacity of ADC on an entrance $\Delta L_{eq} = \log_2\eta$, where η, is determined by formula (10). In this example, the dominant signal was taken at frequency 0. For other cases, the IF circuit is supplemented by a phase changer, which is automatically adjusting IF HX "zero" to a frequency channel numbered "*k*," in which the dominant signal is detected.

For RO, by introducing random factors into the CP process as an addition to the digital copy of radio signals and a digital copy of random perturbations under CMF conditions, the levels of parasitic components in the spectrum of reproduced signals are reduced by 20-30 dB.

Consequently, digital multichannel frequency filtering (DMF) of the signals in the DRFM device itself and the randomization of rough samples are an alternative to traditional construction, but with multi-bit processing.

6. CONCLUSION

The article explores the possibility of using "low-discharge" ("Rough") readings (statistics) in radio frequency storage technology (DRFM technologies).

It is proposed to use randomization of measurements by the Monte Carlo method, which uses "rough" ("Boolean") statistics in statistical tests to reduce ("smooth") quantization noise.

In order to reduce the requirements for dynamic range in multi-signal mode, it is recommended to identify the dominant signal and exclude it from further processing.

The obtained recommendations are useful for developers of DRFM technology, and the knowledge gained is needed for specialists in the field of randomization of physical measurements.

REFERENCES

1. Lowenschuss O. Coherent RF Memory - New Signal Processing Tool. *IEENEACON-80, Proc. Dayton*. P. 1188-1194 (No. 81/43813 in the GPNTB Fund, Moscow, Russia).
2. Lowenschuss O, Bruce E. Gordon Digital memory system. *US Patent №4280219*, Jul.21.1981.
3. Walter Larry. Development of New Aviation Equipment. *Electronics,* 1986, 10:34.
4. Karmanov YT, Rukavishnikov VM, Shunyayev MI. Study of parameters of digital devices of storage and reproduction of radio signals. *Coll. of sci. papers of Inst. of Eng. and Techn., pp. 70-76*. Chelyabinsk, ChIET Publ., 1980.
5. Karmanov YT, Rodionov VV. Digital processing of radar signals against the background of active interference. *Abstracts of the STC of young scientists of the Ural zone,* p. 15-17. Sverdlovsk, Ural St.Univ. Publ.,1974.
6. Karmanov YuT, Shunyayev MI, Rukavishnikov VM, Habin VA. Digital Generator of Multiple Response Radio

Signals. *Author 's Certificate of the USSR № 187159*, with priority from 5.11.80.

7. Van Brant LB. *Handbook on Methods of Electronic Suppression and Interference Protection with Radar Control.* USA, 1978.
8. 1879BM3 Chip Datasheet (Technical specification) (DSM) Version 1.1. UFKV 431268001 T01K. *Mikroelektronika of STC Modul,* Moscow.
9. Karmanov YT. Problems and Prospects of Development about Digital Devices for Storage and Reproduction of Radio Signals. *Digital Radio Engineering Systems,* 2002-2004, № 5. Chelyabinsk.
10. Gorbachev YN. Limit Capabilities of Randomized Digital Coherent Filtering Procedures. *NIIEIR Funds: MRS, TTE, Ser. ER, № 35, 5 p*. Moscow, VIMI Publ., 1982.
11. Gorbachev YN. Digital methods of ranging in pulse survey radar. *Avtometriya*, 1988, 2:30-35 (in Russ.).
12. Gorbachev YN. About the possibility of reducing the number of quantization levels in digital filters of SDC by using randomized algorithms. *Radiotekhnika*, 1983, 6:45-47 (in Russ.).
13. Gorbachev YN. *Digital processing of radar signals in conditions of coarse (low-discharge) quantization*. Moscow, CNIRTI Publ., 2007, 87 p.
14. Gorbachev YN, Kulikov GV, Spak AV. *Radar: stochastic approach*. Moscow, Goryachaya liniya-Telekom Publ., 2016, 520 p.
15. Gorbachev YN. Stochastic linearization of bearing in adaptive antenna arrays with rough space-time statistics. *Avtomatika i telemekhanika*, 2009, 12:103-114 (in Russ.).
16. Gorbachev YN. Randomization of non-informative parameters of signals in radio channels of communication and location systems: directions of research. *Fizicheskie osnovy priborostroeniya*, 2018, 7(4(30)):24-31. DOI: 10.25210/jfop-1804-024031 (in Russ.).
17. Gorbachev YN. Randomization of reception, processing and generation of signals in radio channels of communication and location systems. *Tsifrovaya obrabotka signalov*, 2017, 4:3-13 (in Russ.).
18. Gorbachev YN. Theorem on stochastic sampling of images in radar and communication. *Zhurnal radioelektroniki* [electronic journal of IRE RAS], 2018, No. 10.
19. Gorbachev YN. Okna v radiolokatsii [Windows in radiolocation]. *Proceedings of the XXI Int. scientific and technical Conference "Radar, Navigation and Communication-RLNC*2015"*, vol. 2, pp. 770-782. Voronezh, VSU Publ., 2015.

DOI: 10.17725/rensit.2020.12.227

Characteristic form of equations of dynamics of media of complex structure

George G. Bulychev

MIREA-Russian Technological University, http://www.mirea.ru/
Moscow 119454, Russian Federation

E-mail: geo-bulychev@mail.ru

Received November 28, 2019, reviewed December 20, 2019, accepted December 27, 2019

Abstract. **The article presents the construction of the characteristic form of the equations of dynamics of elastic media for which the stiffness matrix is asymmetric. For diagonalization of the asymmetric matrix, the procedure of unitary triangularization by Schur and additional transformations are used, which allow to consistently find all parameters of the stress-strain state of the medium. Additional assumptions, such as cosser's pseudocontinuum, are not required. The division of deformations into symmetric and asymmetric parts is not carried out and they are assumed to be small. Then the following information is given about the matrix form of differential invariants, their properties and their usage in problems of dynamics of different media are discussed. An example of their application to the problem of statics is given.**

Keywords: **dynamic processes, spatial characteristics method, numerical modeling, dynamics and destruction of the environments of complex structure**

UDC 539.3

For citation: George G. Bulychev. Characteristic form of equations of dynamics of media of complex structure. *RENSIT*, 2020, 12(2):227-234. DOI: 10.17725/rensit.2020.12.2.227.

Contents

1. INTRODUCTION

Half a century ago, MIPT began the study of grid-characteristic numerical schemes. A direct study of such schemes was conducted under the direction of K.M. Magomedov and A.S. Kholodov [1,2] and, subsequently, their students [3]; The application of these schemes to the problems of mechanics of a deformed rigid body was dealt with by V.N. Kukujanov with students [4,5,6]. This aspect of the problem was considered in more detail in [7]; the procedure for constructing the characteristic form of the equations of dynamics of elastic-viscoplastic bodies with a symmetric stiffness matrix was also considered there. Such a matrix has a large number of materials used in industry and construction, however, there are materials for which this matrix is asymmetric.

These materials include crystals with defects and crystals grown under nonequilibrium conditions [2, 3]. In addition, dynamic dynamics loading of statically stressed structures [4] and loading of thin-walled structures taking into account the influence of shear deformation [5] leads to the equations of dynamics and statics with an asymmetric elastic matrix.

In [1], when constructing the characteristic equations of dynamics, two main methods were used: 1) the construction of the matrix form of the equations of motion and the defining equations, and 2) the diagonalization of symmetric matrices, which are the main

minors of the third order of the stiffness matrix. After that, the final form of the equations was obtained using simple algebraic transformations.

In the case of an asymmetric stiffness matrix, the matrix form of the equations of dynamics is constructed elementarily, but the procedure of diagonalizing the minors using orthonormal matrices is impossible. Therefore, in this case, it is necessary to use a different trick – unitary triangulation according to Shur. As a result, characteristic equations can still be constructed, while they possess some new properties.

2. THE CONSTRUCTION OF THE CHARACTERISTIC FORM OF EQUATIONS OF DYNAMICS FOR BODIES WITH AN ASYMMETRIC ELASTICITY MATRIX

We construct the characteristic equations using the matrix form for writing variables and Cartesian coordinates $\{x_i\}$. The equations of motion have the form

$$\partial_i p_{ij} = \rho \partial_i V_j; i, j = 1,2,3, \tag{1}$$

where $\partial_i \equiv \partial/\partial x_i$, $\partial t = \partial/\partial t$, p_{ij} are the stresses, V_j – particle velocities, ρ – body material density: summation is carried out here and thereafter on repeated Roman indices.

We introduce additional row matrices

$$e_i = \|\delta_{1i}, \delta_{2i}, \delta_{3i}\|,$$
$$q_{ij} = \left\| \begin{matrix} \delta_{1i}\delta_{1j}, \delta_{2i}\delta_{2j}, \delta_{3i}\delta_{3j}, \delta_{1i}\delta_{2j}, \\ \delta_{1i}\delta_{3j}, \delta_{2i}\delta_{1j}, \delta_{2i}\delta_{3j}, \delta_{3i}\delta_{1j}, \delta_{3i}\delta_{3j} \end{matrix} \right\|, \tag{2}$$

where δ_{ij} is the Kronecker symbol and also write the defined variables: and in the form of matrix rows:

$$V = \|V_1,\ V_2, V_3\|,$$
$$P = \|p_{11},\ p_{22}, p_{33}, p_{12}, p_{13}, p_{21}, p_{23}, p_{31}, p_{32}\|. \tag{3}$$

Using matrices (2) and (3), equation (1) can be written as

$$Q_j \partial_j P^T = \rho \partial_t V^T, \tag{4}$$

somewhere $Q_j = e_i^T q_{ij}$, there

$$Q_1 = \left\| \begin{matrix} 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0 \end{matrix} \right\|,$$

$$Q_2 = \left\| \begin{matrix} 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0 \\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0 \end{matrix} \right\|,$$

$$Q_3 = \left\| \begin{matrix} 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1 \\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \end{matrix} \right\|.$$

We do not consider plasticity; therefore, the relationship between stresses and asymmetric deformations has the form

$$p_{ij} = c_{ijkl}\gamma_{kl}, \tag{5}$$

where $\gamma_{kl} = \partial u_k/\partial x_l$ is the strain tensor, u_k displacement, c_{ijkl} stiffness tensor, which in matrix form has the form of a positive definite asymmetric matrix

$$C = \left\| \begin{matrix} c_{11}\ c_{12}\ c_{13}\ c_{14}\ c_{15}\ c_{16}\ c_{17}\ c_{18}\ c_{19} \\ c_{21}\ c_{22}\ c_{23}\ c_{24}\ c_{25}\ c_{26}\ c_{27}\ c_{28}\ c_{29} \\ c_{31}\ c_{32}\ c_{33}\ c_{34}\ c_{35}\ c_{36}\ c_{37}\ c_{38}\ c_{39} \\ c_{41}\ c_{42}\ c_{43}\ c_{44}\ c_{45}\ c_{46}\ c_{47}\ c_{48}\ c_{49} \\ c_{51}\ c_{52}\ c_{53}\ c_{54}\ c_{55}\ c_{56}\ c_{57}\ c_{58}\ c_{59} \\ c_{61}\ c_{62}\ c_{63}\ c_{64}\ c_{65}\ c_{66}\ c_{67}\ c_{68}\ c_{69} \\ c_{71}\ c_{72}\ c_{73}\ c_{74}\ c_{75}\ c_{76}\ c_{77}\ c_{78}\ c_{79} \\ c_{81}\ c_{82}\ c_{83}\ c_{84}\ c_{85}\ c_{86}\ c_{87}\ c_{88}\ c_{89} \\ c_{91}\ c_{92}\ c_{93}\ c_{94}\ c_{95}\ c_{96}\ c_{97}\ c_{98}\ c_{99} \end{matrix} \right\|.$$

Taking into account the expression of distortions through displacements, as well as using matrices (2), equations (5) can be written in matrix form

$$P^T = CQ_j^T \partial_j u^T \tag{6}$$

or, differentiating by t

$$\partial_t P^T = CQ_j^T \partial_j V^T. \tag{7}$$

Equations (4) and (7) are a complete system of matrix equations for determining matrix rows P and V.

Before performing further transformations, we note a number of properties of the constructed matrices

$$Q_i Q_j^T = \delta_{ij} diag(1,1,1),$$
$$Q_i^T Q_i = diag(1,1,1,1,1,1,1,1,1), \quad (8)$$
$$Q_i C Q_j^T \equiv \theta \Delta \theta^T,$$

where θ – is the orthonormal matrix of the third order, Δ – the matrix is triangular; when $i = j$ the matrix Δ has positive eigenvalues and, if all elements of the central minor of the third order $Q_i C Q_j$ are positive, the largest eigenvalue of the matrix Δ is one [4].

Let the stress wave propagate along the axis x_a of the coordinate system $\{x_i\}$. We multiply (7) on the left by Q_a and apply the Schur theorem to the matrix $Q_\alpha C Q_\alpha^T$, that is, replace it with $\theta_\alpha \Delta_\alpha \theta_\alpha^T$; grouping in (4) and (7) we get

$$\partial_\alpha Q_\alpha P^T - \rho \partial_t V^T = \partial_j Q_j P^T,$$
$$j = \beta, \gamma; \quad \alpha \neq \beta \neq \gamma \neq \alpha; \quad (9)$$
$$\partial_t Q_\alpha P^T - \theta_\alpha \Delta_\alpha \theta_\alpha^T \partial_\alpha V^T = Q_\alpha C Q_j^T \partial_j V^T.$$

We multiply both equations (9) from the left by θ_α^T, introduce the notation:

$$P_\alpha = \theta_\alpha^T Q_\alpha P^T,\ P_\beta = \theta_\alpha^T Q_\beta P^T,$$
$$P_\gamma = \theta_\alpha^T Q_\gamma P^T,\ P_\gamma = \theta_\alpha^T Q_\gamma P^T,$$
$$\Psi_{\alpha\beta} = \theta_\alpha^T Q_\alpha C Q_\beta^T \theta_\alpha,\ \Psi_{\alpha\gamma} = \theta_\alpha^T Q_\alpha C Q_\gamma^T \theta_\alpha,$$

with the help of which equations (9) take the form

$$\partial_\alpha P_\alpha^T - \rho \partial_t V_\alpha^T = \partial_\beta P_\beta^T + \partial_\gamma P_\gamma^T,$$
$$\partial_t P_\alpha^T - \Delta_\alpha \partial_\alpha V_\alpha^T = \Psi_{\alpha\beta} \partial_\beta V_\alpha^T + \Psi_{\alpha\gamma} \partial_\gamma V_\alpha^T. \quad (10)$$

We write the triangular matrix Δ in the form

$$\Delta_\alpha = \begin{Vmatrix} \lambda_\alpha & 0 & 0 \\ \chi_{\alpha\beta} & \lambda_\beta & 0 \\ \chi_{\alpha\gamma} & \chi_{\beta\gamma} & \lambda_\gamma \end{Vmatrix}, \ \lambda_\alpha > \lambda_\beta \geq \lambda_\gamma, \quad (11)$$

and divide it into two parts $\Delta_a = \Lambda_a + E_a$, where Λ_a – is the diagonal matrix

$$\Lambda_\alpha = \begin{Vmatrix} \lambda_\alpha & 0 & 0 \\ 0 & \lambda_\beta & 0 \\ 0 & 0 & \lambda_\gamma \end{Vmatrix}, \quad E_\alpha = \begin{Vmatrix} 0 & 0 & 0 \\ \chi_{\alpha\beta} & 0 & 0 \\ \chi_{\alpha\gamma} & \chi_{\beta\gamma} & 0 \end{Vmatrix}. \quad (12)$$

Next, we write the matrix Λ_a in the form $\Lambda_\alpha = \rho D_\alpha^2,$, where the matrix $D_\alpha = \pm\sqrt{\Lambda_\alpha / \rho} = diag(\pm D_{\alpha\alpha}, \pm D_{\alpha\beta}, \pm D_{\alpha\gamma})$ – is the matrix of longitudinal and transverse velocities of stress waves propagating in both directions along the axis x_a. We multiply the first equation (10) on the left by $|D_a|$ and add the resulting equations, highlighting the characteristic part on the left.

$$(\partial_t + |D_\alpha| \partial_\alpha)(P_\alpha^T - \rho |D_\alpha| V_\alpha^T) =$$
$$= \partial_\beta (\Psi_{\alpha\beta} V_\alpha^T + |D_\alpha| P_\beta^T) + \quad (13)$$
$$+ \partial_\gamma (\Psi_{\alpha\gamma} V_\alpha^T + |D_\alpha| P_\gamma^T) + \boxed{E_\alpha \partial_\alpha V_\alpha^T}.$$

Equations (13) are not characteristic due to the term taken in the frame and containing the derivative along the direction of wave motion. To exclude it, we return to the second equation (10), multiply both its parts on the left by Δ^{-1} and select $\partial_\alpha V_\alpha^T$. In this case, we obtain the equation

$$\partial_\alpha V_\alpha^T = \Delta_\alpha^{-1} (\partial_t P_\alpha^T - \Psi_{\alpha\beta} \partial_\beta V_\alpha^T - \Psi_{\alpha\gamma} \partial_\gamma V_\alpha^T), \quad (14)$$

where:

$$\varepsilon_{\alpha\beta} = -\chi_{\alpha\beta} / \lambda_\alpha \lambda_\beta,\ \varepsilon_{\alpha\gamma} =$$
$$= (\chi_{\alpha\beta} \chi_{\beta\gamma} - \lambda_\beta \chi_{\alpha\gamma}) / \lambda_\alpha \lambda_\beta \lambda_\gamma,$$
$$\varepsilon_{\beta\gamma} = -\chi_{\beta\gamma} / \lambda_\beta \lambda_\gamma,$$
$$\Delta_\alpha^{-1} = \begin{Vmatrix} \lambda_\alpha^{-1} & 0 & 0 \\ \varepsilon_{\alpha\beta} & \lambda_\beta^{-1} & 0 \\ \varepsilon_{\alpha\gamma} & \varepsilon_{\beta\gamma} & \lambda_\gamma^{-1} \end{Vmatrix} = \begin{Vmatrix} \lambda_\alpha^{-1} & 0 & 0 \\ 0 & \lambda_\beta^{-1} & 0 \\ 0 & 0 & \lambda_\gamma^{-1} \end{Vmatrix} + .$$
$$+ \begin{Vmatrix} 0 & 0 & 0 \\ \varepsilon_{\alpha\beta} & 0 & 0 \\ \varepsilon_{\alpha\gamma} & \varepsilon_{\beta\gamma} & 0 \end{Vmatrix} = \Lambda_\alpha^{-1} + E_\alpha^*.$$

We substitute (14) into (13) and carry out the corresponding grouping; as a result, we obtain the equation.

$$(\partial_t + |D_\alpha| \partial_\alpha)(P_\alpha^T - \rho |D_\alpha| V_\alpha^T) =$$
$$= E_\alpha (\Lambda_\alpha^{-1} + E_\alpha^*) \partial_t P_\alpha^T +$$
$$+ \partial_\beta \{[I_3 - E_\alpha (\Lambda_\alpha^{-1} + E_\alpha^*)] \Psi_{\alpha\beta} V_\alpha^T + |D_\alpha| P_\beta^T\} + \quad (15)$$
$$+ \partial_\gamma \{[I_3 - E_\alpha (\Lambda_\alpha^{-1} + E_\alpha^*)] \Psi_{\alpha\gamma} V_\alpha^T + |D_\alpha| P_\gamma^T\},$$

where:

$$E_\alpha(\Lambda_\alpha^{-1}+E_\alpha^*)=\begin{Vmatrix} 0 & 0 & 0 \\ \varepsilon_{\alpha\beta}\lambda_\beta & 0 & 0 \\ -\varepsilon_{\alpha\gamma}\lambda_\gamma & -\varepsilon_{\beta\gamma}\lambda_\gamma & 0 \end{Vmatrix},$$

$I_3 = diag(1,1,1).$

We write (15) in scalar form using the two-valued definition D_a and the following notation:

$$P_\alpha = \|p_{\alpha\alpha}, p_{\alpha\beta}, p_{\alpha\gamma}\|, V_\alpha = \|V_{\alpha\alpha}, V_{\alpha\beta}, V_{\alpha\gamma}\|,$$
$$[I_3 - E_\alpha(\Lambda_\alpha^{-1}+E_\alpha^*)]\Psi_{\alpha\beta}V_\alpha^T \pm D_\alpha P_\beta^T =$$
$$= \Pi_\beta^\pm = \|\Pi_{\alpha\beta}^\pm, \Pi_{\beta\beta}^\pm, \Pi_{\gamma\beta}^\pm\|,$$
$$[I_3 - E_\alpha(\Lambda_\alpha^{-1}+E_\alpha^*)]\Psi_{\alpha\gamma}V_\alpha^T \pm D_\alpha P_\gamma^T =$$
$$= \Pi_\gamma^\pm = \|\Pi_{\alpha\gamma}^\pm, \Pi_{\beta\gamma}^\pm, \Pi_{\gamma\gamma}^\pm\|,$$

with which (15) can be represented in the form of three pairs of scalar equations

$$(\partial_t \pm D_{\alpha\alpha}\partial_\alpha)(p_{\alpha\alpha} \mp \rho D_{\alpha\alpha}V_{\alpha\alpha}) = \partial_\beta\Pi_{\alpha\beta}^\pm + \partial_\gamma\Pi_{\alpha\gamma}^\pm,$$
$$(\partial_t \pm D_{\alpha\beta}\partial_\alpha)(p_{\alpha\beta} \mp \rho D_{\alpha\beta}V_{\alpha\beta}) =$$
$$= \varepsilon_{\alpha\beta}\lambda_\beta\partial_t p_{\alpha\alpha} + \partial_\beta\Pi_{\beta\beta}^\pm + \partial_\gamma\Pi_{\beta\gamma}^\pm, \quad (16)$$
$$(\partial_t \pm D_{\alpha\gamma}\partial_\alpha)(p_{\alpha\gamma} \mp \rho D_{\alpha\gamma}V_{\alpha\gamma}) =$$
$$= -\varepsilon_{\alpha\gamma}\lambda_\gamma\partial_t p_{\alpha\alpha} - \varepsilon_{\beta\gamma}\lambda_\gamma\partial_t p_{\alpha\beta} + \partial_\beta\Pi_{\gamma\beta}^\pm + \partial_\gamma\Pi_{\gamma\gamma}^\pm,$$

of which we successively find the voltage – particle velocity pairs: from the first pair of equations p_{aa} and V_{aa}, from the second pair of equations, substituting the previously obtained – p_{aa} and V_{aa} into the right-hand sides, next pair – $p_{a\beta}$ and $V_{a\beta}$, finally, from the third pair, substituting into the right-hand sides all the previously obtained voltages and particle velocities, obtain remaining $p_{a\gamma}$ and $V_{a\gamma}$. After that, it is possible to reconstruct the row matrices P_a and V_a and use them further to determine the remaining stresses P_β and P_γ at fixed discontinuities. As you know, in this type of equations, the spatial derivative of the defined variables in the direction of wave motion (here it is $\partial_\alpha V_\alpha^T$) should be absent.

We return to equation (7), multiply both its parts on the left by $\theta_\alpha Q_\beta$, express $\partial_\alpha V_\alpha^T$ using (14) and using the notation (10) we obtain for P_β^T

$$\partial_t(P_\beta^T - \theta_\alpha^T Q_\beta C Q_\alpha \theta_\alpha \Delta_\alpha^{-1} P_\alpha^T) =$$
$$\theta_\alpha^T Q_\beta C[I_3 - Q_\alpha^T \theta_\alpha \Delta_\alpha^{-1} \theta_\alpha^T Q_\alpha C](Q_\beta^T \theta_\alpha \partial_\beta V_\alpha^T + \quad (17)$$
$$+ Q_\gamma^T \theta_\alpha \partial_\gamma V_\alpha^T).$$

Similarly, multiplying (7) from the left by $\theta_\alpha Q_\gamma$ for P_γ^T, we obtain

$$\partial_t(P_\gamma^T - \theta_\alpha^T Q_\gamma C Q_\alpha \theta_\alpha \Delta_\alpha^{-1} P_\alpha^T) =$$
$$= \theta_\alpha^T Q_\gamma C[I_3 - Q_\alpha^T \theta_\alpha \Delta_\alpha^{-1} \theta_\alpha^T Q_\alpha C](Q_\beta^T \theta_\alpha \partial_\beta V_\alpha^T + \quad (18)$$
$$+ Q_\gamma^T \theta_\alpha \partial_\gamma V_\alpha^T).$$

Having determined P_α^T, P_β^T, P_γ^T and V_α^T, the initial stresses and particle velocities are found by the formulas

$$P^T = Q_\alpha^T \theta_\alpha P_\alpha^T + Q_\beta^T \theta_\alpha P_\beta^T + Q_\gamma^T \theta_\alpha P_\gamma^T,$$
$$V^T = \theta_\alpha V_\alpha^T. \quad (19)$$

This completes the construction of the characteristic form of the equations of dynamics of media with an asymmetric stiffness matrix. Naturally, in practically important cases, this problem should be solved by numerical methods. However, it should be noted that the matrix approach used in this paper reduces the determination of the eigenvalues of matrices to third-order algebraic equations, which allows one to analytically perform the following operations:

1) find all the elements of the lower triangular matrices Λ_a for all directions of wave propagation $\{x_i\}$ and, therefore, the matrices Λ_a and E_a;

2) find the matrices θ_a corresponding to them

3) find all the coupling coefficients $\varepsilon_{\alpha\beta}$, $\varepsilon_{\alpha\gamma}$, $\varepsilon_{\beta\gamma}$ of the inverse matrix Δ_α^{-1} and then the matrices E_α^*, $E_\alpha(\Lambda_\alpha^{-1}+E_\alpha^*)$, and all unchanged matrices in equations (16)-(18).

After these operations, the machine account for solving applied problems is sharply reduced.

3. SOME PROPERTIES OF CONSTRUCTED MATRICES AND MATRIX DIFFERENTIAL INVARIANTS

Analyzing the above construction of the characteristic equations, as well as the construction carried out in [1], it is easy to see that each uses one kind of auxiliary matrices. In [1] these are matrices $R_i = e_j^T r_{ij}$, where

$$r_{ij} = \left\| \begin{matrix} \delta_{1i}\delta_{1j}, \delta_{2i}\delta_{2j}, \delta_{3i}\delta_{3j}, \delta_{1i}\delta_{2j} + \\ +\delta_{2i}\delta_{1j}, \delta_{1i}\delta_{3j} + \delta_{3i}\delta_{1j}, \delta_{2i}\delta_{3j} + \delta_{3i}\delta_{2j} \end{matrix} \right\|$$

In this paper, these are matrices $Q_i = e_j^T q_{ij}$. The structure of these matrices is simple – they are dyadic products of unit one-line three-component matrices e_j, isolating the components of the vector, also written as a one-line matrix, and similarly written unit matrices r_{ij} and q_{ij}, isolating the components of symmetric and asymmetric tensors, written in the same form.

An important feature of these matrices is that only one matrix is used in each of the considered processes of medium motion.

So the movement of a symmetrically elastic medium can be represented by the following cycle – **Fig. 1**, where the main role in topology is played by matrices R_i: in the equations of motion they lower the dimension (row matrices Σ), and in the defining equations increase the dimension (row matrices V).

Matrices e_j, R_i and Q_i form the corresponding bases in three, six and nine-dimensional spaces, i.e.

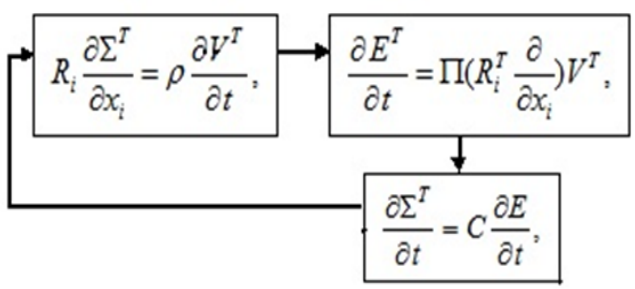


Fig. 1. *The cycle of motion of a symmetrically elastic medium.*

Table 1

Matrix differential invariants

$div\,\vec{v} = e_i \partial_i V^T$	$grad\,\varphi = e_i^T \partial_i \varphi$
$div\,\hat{\sigma} = R_i \partial_i \Sigma^T$	$def\,\vec{v} = \Pi(R_i^T \partial_i V^T)$
$div\,\hat{p} = Q_i \partial_i P^T$	$Grad\,\vec{v} = Q_i^T \partial_i V^T$
$rot\,\vec{v} = S_i \partial_i V^T$, $S_i = \in_{ijk} e_j^T e_k$, $\in_{ijk}$ $-V^T$ – Levi - Civita tensor	

$$\begin{aligned} &e_i^T e_i = I_3,\ R_i^T R_i = \Pi^{-1} I_6, \\ &Q_i^T Q_i = I_9;\ \Pi^{-1} = diag(1,1,1,2,2,2), \end{aligned} \qquad (20)$$

where I_3, I_6 and I_9 are the unit matrices of the third, sixth and ninth order, which allows, using only their pairs, to find all the components of the velocity vector and stress tensors in the problems under consideration.

Table 1 presents the matrix differential invariants encountered in the problems of continuum dynamics and field theory and written in the Cartesian coordinate system $\{x_i\}$. Here is indicated: $\vec{v}$, $\hat{\sigma}$ and $\hat{p}$ – a vector, symmetric and asymmetric tensors, V, Σ and P – they are also written in the form of single-line matrices; φ – scalar, $\partial_i = \partial/\partial x_i$.

The matrix S_i corresponds to the vector product of vectors and reflects the orthogonality of the movement and the action caused by it.

In an arbitrary orthogonal coordinate system $\{q^i\}$, the same differential invariants have the form presented in **Table 2**
where

Table 2

Matrix differential invariants
in an arbitrary orthogonal coordinate system

$div\,\vec{v} = [\partial_i + (\partial_i\, ln\,(g_0/h_i))]e_i V^T$	$grad\,\varphi = e_i^T \partial_i \varphi$
$div\,\hat{\sigma} = Div\,\bar{\sigma} = [\partial_i + (\partial_i\, ln\,(g_0/h_i)) + + \in_{ijk} S_j (\partial_k\, ln\, h_i)] R_i \Sigma^T$;	$def\,\vec{v} = \Pi R_i^T [\partial_i + \in_{ijk} S_j (\partial_k\, ln\, h_i)] V^T$
$div\,\hat{p} = Div\,\bar{p} = [\partial_i + (\partial_i\, ln\,(g_0/h_i)) + + \in_{ijk} S_j (\partial_k\, ln\, h_i)] Q_i P^T$;	$Grad\,\vec{v} = Q_i^T [\partial_i + \in_{ijk} S_j (\partial_k\, ln\, h_i)] V^T$
$rot\,\vec{v} = [\partial_k + (\partial_k ln\, h_i)] \in_{ijk} e_j^T e_i V^T$	

$$\partial_i = \partial / \partial x_i; \partial x_i = h_i \partial a^i; h_i = |\, grad\ a^i\,|;$$
$$g_0 = h_1 h_2 h_3; i, j, k, \alpha, \beta = 1, 2, 3;\ i \neq j \neq k \neq i.$$

The expressions for matrix differential invariants presented in Tables 1 and 2 are convenient for matrix transformations: they have a block form, are written in a Cartesian (or accompanying Cartesian) coordinate system, and contain matrix rows of physical components of the defined variables that are affected by scalar - differential – ∂_i and algebraic – $(\partial_i \ln(g_0/h_i))$ and – $(d_k \ln h_i)$ operators. These equations contain four types of matrices, – e_i, R_i, Q_i and S, "stitching" individual scalar equations into groups and determining the structure of these groups.

Like the original symbolic ones, the constructed matrix relations are invariant under coordinate transformations, however, the use of these properties in matrix relations has its own peculiarities. Consider the transformation of matrix invariants when rotating the coordinate axes.

We introduce the Cartesian coordinate system $\{y_i\}$, with the axes rotated relative to the axes $\{x_i\}$ and the center coinciding with the center of the system $\{x_i\}$, and the transition matrix $\Theta_3 = e_j^T \theta_{ji} e_i$ from coordinates $\{x_i\}$ to coordinates $\{y_i\}$; obviously $\theta_{ji} = \partial y_i / \partial x_i$.

We transform the invariants written in tables 1 and 2 to variables $\{y_i\}$, for which we introduce the matrices M_3, M_6 and M_9, defined using the relations

$$\Theta_3 \theta_{ji} S_i = S_j M_3, \Theta_3 \theta_{ji} R_i = \\ = R_j M_6, \Theta_3 \theta_{ji} Q_i = Q_j M_9. \quad (21)$$

Multiplying expressions (21) from the left, respectively, by S_j^T, R_j^T and Q_j^T, and using equalities (20), we define these matrices in the form

$$M_3 = 0.5 S_j^T \Theta_3 \theta_{ji} S_i,\ M_6 = \Pi R_j^T \Theta_3 \theta_{ji} R_i, \\ M_9 = Q_j^T \Theta_3 \theta_{ji} Q_i. \quad (22)$$

It is easy to show that both $M_3 = \Theta_3$ and $M_9 = \Theta_9$ are orthonormalized and are rotation matrices in three-dimensional and nine-dimensional space, the matrix M_6 provides rotation of the components of the symmetric rotation tensor in six-dimensional space, but is not orthonormal, $|M_6| = 1$.

Using the properties of the constructed matrices e_i, R_i, S_i and Q, and carrying out a series of identical transformations, it is easy to construct formulas that allow one to recalculate the matrix differential invariants written in the concomitant (lower index x) coordinate system rotated relative to it (lower index y)

$$div_x V^T = div_y \Theta_3 V^T, \quad \Theta_3 grad_x T = grad_y \Theta_3 T; \\ \Theta_3 rot_x V^T = rot_y \Theta_3 V^T, \quad \Theta_3 Div_x \Sigma^T = Div_y \Theta_6 \Sigma^T; \\ \Theta_3 Div_x P^T = Div_y \Theta_9 P^T, \quad \Theta_6 \Pi^{-1} def_x V^T = \Pi^{-1} def_y \Theta_3 V^T; \quad (23)$$

$$Grad_x V^T = \Theta_9^T\, Grad_y\, \Theta_3 V^T,$$

or, in general terms,

$$inv_x \Psi^T = \Theta_i^T\, inv_y\, \Theta_j \Psi^T. \quad (24)$$

Comparing formulas (23)-(24), we can formulate the following statement:

To write the matrix form of any differential invariant (*inv*) from a column matrix Ψ^T in a Cartesian coordinate system $\{y_i\}$, rotated relative to the accompanying Cartesian system $\{x_i\}$ by a certain angle, specified using the rotation matrix Θ_3, it is necessary:
1) multiply the transformed matrix invariant on the left by the transposed rotation matrix Θ_i, the order of which coincides with the number of equations of the group; for deformation should be used as $\Pi^{-1} def$; $\Theta_1 \equiv 1$;
2) replace it $\partial / \partial x_i$ with $\partial / \partial y$;
3) in his argument, replace Ψ^T with the product $\Theta_j \Psi^T$, where the matrix Θ_j has an order that matches the number of elements in the column matrix Ψ^T.

Consider an example where auxiliary matrices, matrix differential invariants, and their properties are useful in solving the problem.

Let in the defining equation the elastic body statics $\hat{\sigma} = c\,def\,\vec{u}$, where $c = c^{ijkl}\,\vartheta_i\vartheta_j\vartheta_k\vartheta_l$, and c^{ijkl}– are the contravariant components of the tensor corresponding to the symmetric nondegenerate stiffness matrix. Let at the boundary with the normal x_a, all derivatives of the displacement vector $\vec{u}$ except $\partial/\partial x_a$, and three components of the stress tensor $\hat{\sigma}$ (σ_{aa}, $\sigma_{a\beta}$, $\sigma_{a\gamma}$, $a \neq \beta \neq \gamma$) on this boundary are known.

It is necessary to construct the simplest equations for determining the remaining components of the stress tensor.

Decision. We write the original equation in matrix form by introducing a non-degenerate matrix $C=\Pi r_{ij}^T c^{ijkl} r_{kl}\Pi$, which is the matrix expression of c, both the row matrix of stresses – Σ and displacements – U.

We multiply the resulting equation on the left by C^{-1}, and we get the equation

$$C^{-1}\Sigma^T = \Pi^{-1} def_x U^T. \tag{25}$$

We draw attention to the fact that in the operator def_x the derivatives $\partial/\partial x_a$ of the row matrix U are multiplied by the matrix R_α^T, therefore, to exclude them from equation (25), it is enough to multiply the latter by the matrix orthogonal R_a.

We construct such a matrix in the form $N_\alpha = I_6 \ominus R_\alpha$, where $\ominus$ – is the operation of deleting rows corresponding to the matrix R_a from the matrix I_6. The matrices R_a and N_a together form a six-dimensional basis (of two three-dimensional), obviously

$$\begin{aligned} &N_\alpha R_\alpha^T = 0,\ R_\alpha R_\alpha^T = I_3, \\ &N_\alpha N_\alpha^T = I_3,\ R_\alpha^T R_\alpha + N_\alpha^T N_\alpha = I_6. \end{aligned} \tag{26}$$

We divide the components of the stress matrix row Σ into two groups: the stresses acting on the site with a normal x_a, – $\Sigma_\alpha^T = R_\alpha \Sigma^T$ and all the others – $\Sigma_n^T = N_\alpha \Sigma^T$, and, obviously, $\Sigma^T \equiv R_\alpha^T \Sigma_\alpha^T + N_\alpha^T \Sigma_n^T$. Let us $A \equiv C^{-1}, A_{nn} \equiv N_\alpha A N_\alpha^T, A_{nr} \equiv N_\alpha A R_\alpha^T$ then write equation (16) in the form

$$A_{nn}\Sigma_n^T + A_{nr}\Sigma_\alpha^T = N_\alpha \Pi^{-1} def_x U^T. \tag{27}$$

The matrix A_{nn} is the main submatrix of a non-degenerate matrix A and, therefore, is non-degenerate. Therefore, equation (27) can be divided on the left by A_{nn}, after which it will take the form

$$\Sigma_n^T = A_{nn}^{-1}(N_\alpha \Pi^{-1} def_x U^T - A_{nr}\Sigma_\alpha^T). \tag{28}$$

Equation (28) satisfies the given requirements, however, when constructing it, it is necessary to calculate the matrix A, the inverse C, the submatrices A_{nn} and A_{nr}, and the matrix A_{nn}^{-1}. To simplify the calculations, we use the identity $AC \equiv I_6$ and relations (26) that determine the decomposition of matrices A and C the basis of interest to us.

For submatrices A_{nn} and A_{nr}, we obtain the following relations:

$$A_{nn}C_{nn} + A_{nr}C_{rn} = I_3,\ A_{nn}C_{nr} + A_{nr}C_{rr} = 0 \tag{29}$$

connecting submatrices of interest with submatrices $\begin{matrix} C_{nn} = N_\alpha C N_\alpha^T,\ C_{nr} = N_\alpha C R_\alpha^T, \\ C_{rn} = R_\alpha C N_\alpha^T,\ C_{rr} = R_\alpha C R_\alpha^T \end{matrix}$ matrices C.

Note that relations (29) *are not a standard block decomposition of matrices A and C*, since the submatrices under consideration do not form continuous blocks in the corresponding matrices, therefore (29) can be considered *as a generalization* of the block decomposition of nondegenerate matrices. For problems with real physical content, this generalization is very significant, since the bases for the decomposition are determined from the context of the problem.

Solving equations (29) we find A_{nn}^{-1} and $A_{nn}^{-1}A_{nr}$, and, substituting these expressions in equation (28), we finally obtain

$$\begin{aligned} &\Sigma_n^T = (C_{nn} - C_{nr}C_{rr}^{-1}C_{rn})N_\alpha \Pi^{-1} def_x U^T + \\ &+C_{nr}C_{rr}^{-1}\Sigma_\alpha^T \end{aligned} \tag{30}$$

equation satisfying all the requirements.

4. CONCLUSION

In the work, the characteristic form of the equations of dynamics of deformable bodies with an asymmetric elastic matrix is constructed. As such bodies, there can be crystals with defects and a number of thin-walled and statically stressed building structures grown under nonequilibrium conditions.

It is shown that, in contrast to bodies with a symmetric matrix, in such bodies, longitudinal and transverse waves interact with each other inside the body. Since it is assumed that the tangential stresses σij and σji are not equal to each other, the obtained equations describe moment media, without additional assumptions, such as, for example, in the Cosserat pseudocontinuum theory [12].

During the construction, the simplest matrices e, R, Q, S were used, which allow both to group the defined variables and to select the necessary components from the constructed groups. They are either an elementary matrix a row of the same dimension as the variable for which they are used, or they are a scalar, vector or dyad product of such row matrices. It turned out that these matrices are enough to construct the matrix form of all differential invariants in both Cartesian and any orthogonal curvilinear coordinate system.

It is convenient to use these matrices for various transformations of coordinate systems, in particular, together with the usual rotation matrix θ, they allow one to construct rotation matrices for three, six, and nine-component vectors and write differential invariants in a rotated coordinate system relative to the original one. In three, six, and nine-dimensional spaces, the matrices e, R, Q, S form unit bases.

The following example shows that such a group approach allows one to quickly and comfortably solve various problems of mathematical physics.

In conclusion, we note that the transformations carried out in [7] and in this work, and the example considered above, do not exhaust all the possibilities of using the mathematical apparatus described above.

REFERENCES

1. Magomedov KM. On the calculation of the desired surfaces in spatial methods of characteristics. *DAN USSR*, 1966, 171(6):1297-1300.
2. Magomedov KM, Colds AC. *Grid-characteristic numerical methods*. Moscow, Nauka Publ., 1988, 288 pp.
3. Petrov IB, Colds AC. On the regularization of discontinuous numerical solutions of hyperbolic equations. *ZhVMiFM*, 1984, 24(8):1172-1188.
4. Kukujanov VN. *Numerical solution of non-one-dimensional problems of the propagation of stress waves in solids*. Vol. 8, 67 p. Moscow, 1967, VC AN USSR.
5. Kukujanov VN, Kondaurov VI. Numerical solution of non-one-dimensional problems of the dynamics of a solid deformable body. In: *Problems of the Dynamics of Elastoplastic Media*. Moscow, Mir Publ., 1975, p. 40-84.
6. Bulychev GG, Kukujanov VN. Dynamic failure of a prestressed fiber composite caused by fiber breakage. *Izvestia RAS, ser. MTT*, 1993, 3:207-214.
7. Georgy G. Bulychev. Method of spatial characteristics in problems of a mechanics of deformable solid body. *Radioelectronics. Nanosystems. Information Technologies* (*RENSIT)*, 2018, 10(1)77-90. DOI: 10.17725/rensit.2018.10.077.
8. Hirt J, Lot I. *Theory of dislocations*. Moscow, Atomizdat Publ., 1972, 598 p.
9. Shafranovsky II. *Crystals of minerals. Curved, skeletal and dendritic forms*. Moscow, Gosgeoltekhizdat Publ., 1961, 332 p.
10. Petrashen GI. *Propagation of waves in anisotropic elastic bodies*. Moscow, Nauka Publ., 1980, p. 225-234.
11. Britvin EI. Formation of a stiffness matrix of thin-walled rods taking into account the influence of shear deformation. *Construction mechanics and calculation of structures*, 2017, 1:23-28.
12. Novatsky V. *Theory of elasticity*. Moscow, Nauka Publ., 1975, p. 797-805.

DOI: 10.17725/rensit.2020.12.235

Calculation of ordering energies by the model potential method taking into account the linear size effect in the Ni-14at.%Pt alloy

Valentin M. Silonov

Lomonosov Moscow State University, http://www.phys.msu.ru/
Moscow 119991, Russian Federation
E-mail: silonov_v@mail.ru

Lkhamsuren Enkhtor

National University of Mongolia, School of Arts and Science, Department of Physics, http://www.num.edu.nm/
Ulaanbaatar 210646, Their Surguuliyin Gudamzh, 1, Mongolia
E-mail: enkhtor@num.edu.mn

Received December 16, 2019, reviewed December 12, 2019, accepted January 14, 2020

Abstract. **For disordered binary solid solutions, a new method is proposed for calculating the ordering energies in an arbitrary coordination sphere, taking into account the linear size effect. Using the model potential of transition metals, the Animalu calculated the ordering energies of the Ni-14at.% Pt solid solution in twelve coordination spheres. The temperature of the order-disorder phase transition was estimated for the Ni-14at.%Pt alloy. Satisfactory agreement with the experimental data is obtained.**

Keywords: **binary alloys, ordering energy, short-range order parameters, pseudopotential method.**

UDC 539.1

Acknowledgments. This work was funded by the Russian-Mongolian grant of the Russian Foundation for Basic Research N 52-44003/19 on the topic "Patterns of formation of the structure and properties of functional composite systems in order to obtain materials for biomedical and radiation-protective purposes".

For citation: Valentin M. Silonov, Lkhamsuren Enkhtor. Calculation of ordering energies by the model potential method taking into account the linear size effect in the Ni-14at.%Pt alloy. *RENSIT*, 2020, 12(2):235-240. DOI: 10.17725/rensit.2020.12.235.

Contents

1. INTRODUCTION

In [1], short-range order in a polycrystalline Ni-14at.%Pt solid solution was studied by diffuse X-ray scattering, and its parameters were determined on the first six coordination spheres with allowance for the size effect. Earlier, in [2], the short-range order was studied in a single-crystal alloy Ni-23.2at.% Pt and the long-range nature of interatomic interactions in disordered solid solutions of the nickel-rich Ni-Pt system was revealed. Such studies are of considerable interest, since, as shown in [3], based on data on short-range order parameters in disordered solid solutions, it is possible to construct their atomic-crystalline structure. In this connection, it is also of interest to develop theoretical methods for determining short-range order parameters in disordered solid solutions on an arbitrary number of coordination spheres. In [4], attempts were made to calculate the ordering energies by the pseudopotential method. However, no ordering energies were calculated on the far coordination spheres. In [5], a theory was developed, which was not previously applied, for calculating the ordering energies in a binary alloy on arbitrary

coordination spheres, taking into account the static displacements of atoms due to the size effect.

The aim of this work is to develop a method for calculating the ordering energies in binary alloys based on [5] in an arbitrary coordination sphere taking into account the linear size effect and to perform such a calculation using the example of a disordered Ni-14at.%Pt solid solution.

2. ACCOUNT OF STATIC DISPLACEMENTS IN THE ELECTRONIC THEORY OF CLOSE ORDER

As shown in [5], the configurational energy of a disordered solid solution in the presence of static atomic displacements due to the size effect can be written as:

$$E_{conf} = A_0 - B_0 + C_A C_B \sum_{i\neq 0} C_i \alpha_i V(r_i), \quad (1)$$

where

$$V(r_i) = \left[A(r_i) - B(r_i) \right].$$

In this expression, the function $A(r_i)$ is the fraction of the ordering energy recorded without taking into account the size effect:

$$A(r_i) = V_1^{AA}(r_i) + V_1^{BB}(r_i) - 2V_1^{AB}(r_i) \quad (2)$$

and $B(r_i)$, taking into account the linear size effect, α_i is the short-range order parameter on the ith coordination sphere of radius r_i, C_i is the coordination number, C_A and C_B are the component concentrations.

The energy of pair interaction of atoms of type A entering into formula (2) is written using the normalized characteristic function $G^{AA}(q)$, which includes the contributions of electrostatic interaction and second-order perturbation theory:

$$V_1^{AA}(r_i) = \frac{2\left(Z_A^*\right)}{\pi} \int_0^\infty G^{AA}(q) \frac{\sin qr_i}{qr_i} dq, \quad (3)$$

$$G^{AA}(q) = e^{-q^2/4\eta} - \frac{\Omega_0^2 q^4}{16\pi^2 \left(Z_A^*\right)^2} \left| w_A^0 \right|^2 \frac{\varepsilon(q)-1}{\varepsilon(q)\left(1-f(q)\right)}. \quad (4)$$

The energies $V_1^{BB}(r_i)$ and $V_1^{AB}(r_i)$ are recorded in a similar way. In these expressions, w_A, w_B and Z_A, Z_B are the form factors of model potentials and valency of the components of the alloy of varieties A and B.

The contribution to the ordering energy due to the linear size effect can also be written using normalized characteristic functions in the form [5]:

$$B(r_i) = \Delta_{AA,i}\left[V_2^{AB}(r_i) - V_2^{AA}(r_i)\right] - \Delta_{BB}\left[V_2^{BB}(r_i) - V_2^{AB}(r_i)\right] - \\ -\Delta_{AA,i}\left[V_1^{AB}(r_i) - V_1^{AA}(r_i)\right] + \Delta_{BB}\left[V_1^{BB}(r_i) - V_1^{AB}(r_i)\right], \quad (5)$$

where

$$V_2^{AA}(r_i) = \frac{2\left(Z_A^*\right)}{\pi} \int G^{AA}(q) \cos(qr_i) dq. \quad (6)$$

The functions $V_2^{BB}(r_i)$ and $V_2^{AB}(r_i)$ are defined similarly.

In this work, in calculating the ordering energies, we used form factors of the model potential of transition metals of Animalu, which have the form [6]:

$$W^{bare}(q) = F\left(\vec{k}_F, \vec{k}_F + \vec{q}\right) + B(q), \quad (7)$$

$$B(q) = -\frac{8\pi A_2}{\Omega_0 q^3}\left[\sin(qR_m) - qR_m \cos(qR_m)\right] - \frac{8\pi Z}{\Omega_0 q^2}\cos(qR_m) + \\ + \left[\frac{4\pi |E_C|}{\Omega_0 q^3} - \frac{24\pi Z \alpha_{eff}}{\Omega_0 q^2 (qR_C)^3}\right] \times \left[\sin(qR_m) - qR_m \cos(qR_m)\right]. \quad (8)$$

For $\left|\vec{k}_F + \vec{q}\right| = \vec{k}_F$

$$F\left(\vec{k}_F, \vec{k}_F + \vec{q}\right) = -\frac{4\pi R_m^3}{\Omega_0} \sum_l (2l+1) \times \\ \times (A_l - C)\left[j_l^2(x) - j_{l-1}(x) j_{l+1}(x)\right] R_l(\cos\theta). \quad (9)$$

For $\left|\vec{k}_F + \vec{q}\right| \neq \vec{k}_F$

$$F\left(\vec{k}_F, \vec{k}_F + \vec{q}\right) = -\frac{8\pi R_m^3}{\Omega_0 \left(x^2 - y^2\right)} \sum_l (2l+1)(A_l - C) \times \\ \times \left[x j_{l+1}(x) j_l(y) - y j_{l+1}(y) j_l(x)\right] P_l(\cos\theta'), \quad (10)$$

where

$$x = k_F R_m,\ y = \left|\vec{k}_F - \vec{q}\right| R_m,\ C = \frac{Z}{R_m},$$

$$\cos\theta = \left(1 - \frac{q^2}{2k_F^2}\right),\ \cos\theta' = \frac{x^2 + y^2 - (qR_m)^2}{2xy},$$

j_l are the Bessel spherical functions, $R_l(\cos\theta)$ are the Legendre polynomials, Ω_0 is the atomic volume,

$$\varepsilon(q) = 1 + \left[1 - f(q)\right]\frac{4\pi Ze^{*2}}{\Omega_0 q^2}\left(\frac{2}{3}E_F\right)^{-1} \times \times\left[\frac{1}{2} + \frac{4k_F^2 - q^2}{8k_F q}\ln\left|\frac{2k_F + q}{2k_F - q}\right|\right], \quad (11)$$

$E_F = \frac{\hbar^2 k_F^2}{2m^*}$, k_F – is the Fermi momentum, m* is the effective mass of the electron, $e^{*2} = (1 + \alpha_{eff})e^2$, and $f(q)$ is the correction for the exchange and correlation of electrons. In the work, the function proposed by Hubbard and Sham was used.

$$f(q) = \frac{q^2}{2\left(q^2 + k_F^2 + k_s^2\right)},\ k_s^2 = \frac{2k_F}{\pi}. \quad (12)$$

The used parameter values of the nickel and platinum pseudopotentials are given in **Table 1**.

The temperature correction was introduced by multiplying the form factors of nickel and platinum by the factors exp(-M), where

$$M = \frac{q^2}{2}\left\langle u_s^2\right\rangle = \frac{q^2}{2}\frac{3\hbar^2 T}{Mk_B\Theta_D^2}\left(\Phi(X) + \frac{X}{4}\right), \quad (13)$$

$\left\langle u_s^2\right\rangle$ are the mean square displacements, $\hbar$ is the Planck constant, T is the temperature, M is the mass of the atom, k_B is the Boltzmann constant, Θ_D is the Debye temperature, $\Phi(X)$ is the Debye function, $X = \Theta_D/T$. The calculations were carried out for a temperature of 1000°C.

Table 1

The values of the nickel and platinum pseudopotentials

	A_0	A_1	A_2	R_m	Θ	Z	m*	R_C	α_{eff}	$\lvert E_C\rvert$
Ni	0.99	1.05	0.98	2.2	73.6	2	1.0	1.304	0.063	0.093
Pt	0.97	1.11	0.85	2.6	101.6	2	1.0	1.512	0.071	0.091

3. METHOD FOR CALCULATING THE STATIC DISPLACEMENT PARAMETERS

The parameters of the static displacements $\Delta_{AA,i}$ and $\Delta_{BB,i}$ were calculated using the scheme for changing the sizes of nickel and platinum atoms during the formation of a solid solution, which is shown in **Fig. 1** and is similar to the scheme used in [7]. It is assumed that the dependences of atom sizes on concentration are close to linear and parallel to each other. In Fig. 1, these dependencies are shown by solid lines. Let $r^0_{AA,i}$ and $r^0_{BB,i}$ there be interatomic distances of atoms in pure metals, and $r^1_{AA,i}$ and $r^1_{BB,i}$ – in the alloy. Then you can get:

$$r^1_{AA} = r^0_{AA} + \frac{r^0_{BB} - r^0_{AA}}{K}C_B, \quad (14)$$

$$r^1_{BB} = r^0_{BB} - \frac{r^0_{BB} - r^0_{AA}}{K}C_A. \quad (15)$$

In equations (14) and (15), K is a fitting parameter. In the proposed model, the displacement parameters will be:

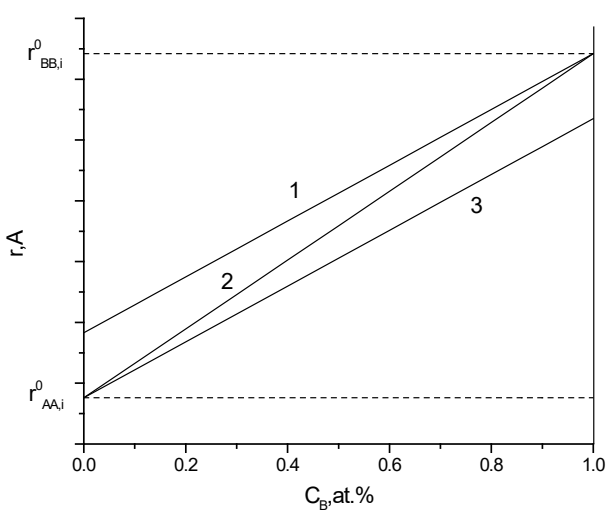


Fig. 1. *Model dependences of interatomic distances between atoms of grades A and B on the concentration of the alloy in a linear approximation. Line 1 is the dependence $r^1_{BB,i}$ on concentration, line 3 is the dependence $r^1_{AA,i}$ on concentration, and line 2 is the dependence of the radius of the ith coordination sphere ri on the alloy concentration. Dotted lines are the distances between atoms in pure metals.*

$$\Delta_{AA,i} = \frac{r^1_{AA,i}}{r_i} - 1 \text{ and } \Delta_{BB,i} = \frac{r^1_{BB,i}}{r_i} - 1. \quad (16)$$

In (16), r_i is the radius of the i-th coordination sphere can be estimated by a linear dependence on concentration (in Fig. 1, it is also shown by a solid line).

In this work, the fitting parameter K was found by fitting the short-range order parameters $\alpha^{\exp[1]}_{lmn}$ of the polycrystalline alloy Ni-14at.%Pt, which was determined from the intensity of diffuse x-ray scattering measured in [1]. Moreover, for a number of K values, the sums of standard deviations $\frac{1}{N}\sum_{i=1}^{N}\left(\alpha_i^{th}C_i - \alpha_i^{\exp}C_i\right)^2$, were calculated where C_i are coordination numbers, and the true value was taken to be the corresponding minimum value. The theoretical values of the Ni-14at.%Pt alloy parameters α_i^{th} were determined by the Krivoglaz–Klepp–Moss method [8] from the ordering energies $V^{th}(r_i)$ calculated by the pseudopotential method. First, the Fourier components of the values $V^{th}(r_i)$ were calculated using the relation

$$\alpha\left(\vec{k}\right) = \frac{1}{1 + 2C_AC_B V\left(\vec{k}\right)/k_B} \quad (17)$$

the Fourier transforms of the short-range order parameters $\alpha\left(\vec{k}\right)$ were found. From these values, the short-range order parameters $\alpha^{th}(r_i)$ were determined using the inverse Fourier transform.

To verify the proposed scheme for calculating the ordering energies, we estimated the value of the temperature of the order-disorder phase transition T_c^{th}, which, according to [9], is related to the value of the Fourier transform of the ordering energy $V\left(\vec{k}_m\right) = \sum_{\vec{r}} V(\vec{r}) e^{i\vec{k}_m\vec{r}}$ by the relation:

$$T_c^{th} = \min\left(-\frac{2C_AC_B}{k_B}V^{th}\left(\vec{k}_m\right)\right), \quad (18)$$

where $\vec{k}_m$ is one of the vectors of the star associated with the order-disorder phase transformation. The calculation was carried out for the reflex (100).

4. CALCULATION RESULTS AND DISCUSSION

During the formation of a solid solution, the sizes of the atoms of the components change [7]. So atoms that were larger in the initial state than in solid solution should decrease in size and fit into the middle lattice of a virtual crystal. In contrast, smaller atoms should increase them. In both cases, the component atoms will be displaced from the nodes of the crystal lattice of the middle crystal, the lattice parameters of which are found from reflections. At the same time, it is important to develop methods for assessing such changes. In this paper, an attempt is made to make such estimates in the simplified model shown in Fig. 1. **Table 2** shows the atomic sizes of Ni and Pt in pure metals and in a solid solution of Ni-14at.% Pt. It is seen that in the solid solution, nickel atoms increase their size by 0.033Å to a value of 2.528Å, and platinum atoms decrease by 0.197Å to a value of 2.771Å. In this case, the decrease in the size of platinum atoms exceeds the increase in the size of nickel atoms by almost six times. It turned out that if in the initial state the sizes of nickel and platinum atoms differed by 0.283 Å, then this difference in the solid solution was 0.046 Å. This largely determined the calculated values of the parameters $\Delta_{NiNi,i}$ and $\Delta_{PtPt,i}$, which turned out to be equal to –0.0028 and 0.0182, respectively. This ratio corresponds to the predominant contribution of static distortions due to the influence of platinum atoms.

Table 3 shows the values of the ordering energy components $A(r_i)$ and $B(r_i)$ for the first twelve

Table 2

The sizes of Ni and Pt atoms in pure metals and solid solution Ni-14at.%Pt

Content, at%Pt	r_i, Å	r_{Ni}, Å	r_{Pt}, Å
0	2.488	2.488	-
14	2.528	2.521	2.674
100	2.771	-	2.771

Table 3
The values of the ordering energies $A(r_i)$ and $B(r_i)$ for the first twelve coordination spheres

Sphere number	$A(r_i)$, meV	$B(r_i)$, meV
1	16.78	-34.22
2	3.17	12.62
3	0.67	-2.35
4	-0.81	-4.28
5	0.15	1.12
6	0.48	2.10
7	0.01	-1.80
8	-0.30	1.34
9	-0.13	1.31
10	0.15	1.90
11	0.19	0.20
12	0.02	-4.96

coordination spheres, calculated using the model of the transitional metal potential of Animalu and taking into account static displacements. It can be seen that, in the first sphere, the energy contribution $B(r_i)$ is twice as large as the component $A(r_i)$, which indicates the importance of taking into account static displacements in estimating the ordering energy of disordered solid solutions. Also from the data Table 3 shows that, in all other coordination spheres, the contribution of energy $B(r_i)$ is predominant. This explains the previously failed attempts of such calculations [4-5].

The results of calculating the ordering energies of the Ni-14at.%Pt alloy for the first twelve coordination spheres are shown in the second column of **Table 4.** They were obtained with a fitting parameter K of 1.233. It is seen that the calculated energies $V^{\text{th}}(r_i)$ corresponding to the minimum of standard deviations $\frac{1}{N}\sum_{i=1}^{N}\left(\alpha_i^{th}C_i - \alpha_i^{\exp}C_i\right)^2$, with increasing coordination sphere number are of a characteristic alternating character. Comparing the calculated and experimental values of the short-range order parameters of the Ni-14at.%Pt alloy given in the third and fourth columns of the Table, we can note their satisfactory agreement. So for most coordination areas they coincided in sign and were close in magnitude. The value $V\left(\vec{k}_m\right)$ calculated using the obtained ordering energies

Table 4
The ordering energies $V^{\text{th}}(r_i)$, theoretical $\alpha^{th}(r_i)$ and experimental $\alpha^{[1]}(r_i)$ values of the short-range order parameters of the Ni-14at.%Pt alloy

i	$V^{\text{th}}(r_i)$, meV	$\alpha^{th}(r_i)$	$\alpha^{[1]}(r_i)$
1	50.40	-0.111	-0.041
2	-9.23	0.082	0.170
3	2.99	0.006	0.000
4	3.40	0.013	0.017
5	-0.95	-0.013	-0.010
6	-1.58	0.016	0.093
7	1.77	-0.009	-0.025
8	-1.62	0.001	0.024
9	-1.42	0.007	0.008
10	-1.72	0.06	0.030
11	-0.01	-0.001	-0.061
12	4.89	-0.002	0.013

$V^{\text{th}}(r_i)$ was −287 meV, and the value of the order – disorder phase transition temperature was 519°C, which is consistent with the data given in [10].

In **Figure 2** shows the dependence of the ordering energies on the first twelve coordination spheres of the Ni-14at.%Pt alloy on the radius of the coordination sphere. For comparison, in Fig. 2 also shows the dependence of the ordering energies of the single-crystal alloy Ni-23.2at.%Pt in the first eight spheres, calculated in

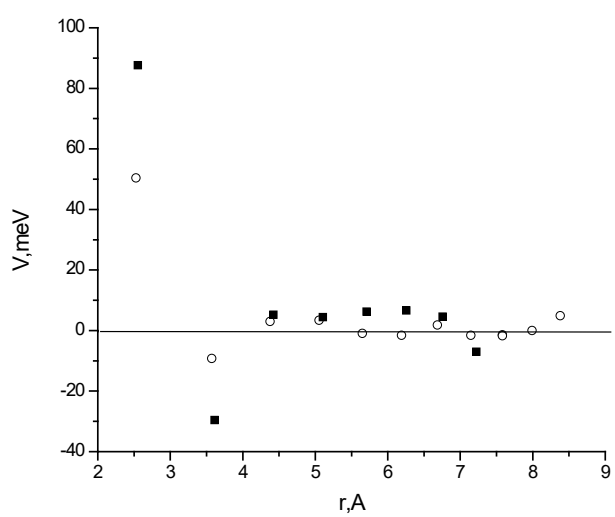


Fig. 2. *The dependences of the ordering energies for the Ni-14at.% Pt and Ni-23.2 at.% Pt alloys on the radii of the coordination spheres:* ○ *– for the Ni-14at.% Pt alloy;* ■ *– for the Ni-23.2 at.% Pt alloy according to* [2].

[2] by the Krivoglaz–Klepp–Moss method using short-range order parameters in the first fifteen coordination spheres, which were experimentally determined by the *X*-ray method. It can be seen from the figure that both dependences have a similar quasi-oscillating character. The values of the ordering energies for the first four and eighth coordination spheres coincided in sign. For the Ni-23.2at.%Pt alloy, the ordering energies in the first two coordination spheres were slightly higher in absolute value than the corresponding values for the Ni-14at.%Pt alloy. In the third and fourth spheres, they turned out to be close in magnitude. In general, their satisfactory coincidence can be noted. This suggests the promise of using pseudopotentials of transition metals [6] for calculating the ordering energies and related characteristics of disordered solid solutions.

5. CONCLUSION

Thus, taking into account the linear size effect in the electronic theory, for the Ni-14at.%Pt alloy, it was possible to carry out a numerical calculation of the ordering energies and the neighbor parameters for twelve coordination spheres. In the calculations, the fitting of the theoretical values of the short-range order parameters to their experimental values determined by the method of diffuse x-ray scattering was used. The contribution of static atomic displacements to the ordering energy is estimated. The obtained values of the ordering energies made it possible to estimate the temperature of the order-disorder phase transition for the Ni-14at.%Pt alloy.

Acknowledgments. *This work was funded by the Russian-Mongolian grant of the Russian Foundation for Basic Research N 52-44003/19 on the topic "Patterns of formation of the structure and properties of functional composite systems in order to obtain materials for biomedical and radiation-protective purposes".*

REFERENCES

1. Enkhtor L, Silonov VM. The short range order and its energy characteristics in the Ni-14at.%Pt alloy. *Vestnik MSU, Ser. 3. Physics. Astronomy,* 2019, 2:73-76.
2. Engelke M, Schonfeld B, Ruban AV. Grazing incidence diffraction and first-principles calculation. *Phys. Rev. B.*, 2010, 81:054205-1-13.
3. Saha DK, Koga K, Ohshima K. Short-range order in Cu-Pd alloys. *J.Phys.: Condens. Mat.*, 1992, 4:10093-10102.
4. Katsnelson AA, Silonov VM, Khwaja FA. Electronic theory of short range in alloys using pseudopotential approximation and its comparison with experiments. *Phys. Stat. Sol. (b)*, 1979, 91:11-33.
5. Katsnelson AA, Mehrabov AOO, Silonov VM. On the contribution to the energy and structural characteristics of ordering calculated by the pseudopotential method. *FMM*, 1976, 42(2):278-283.
6. Animalu AOE. Electronic structure of transition metals. I. Quantum defects and model potentials. *Phys. Rev.,* 1973, 8(8):3542-3554.
7. Flinn PA, Averbach BL, Cohen M. Local atomic arrangements in gold-copper alloy. *Acta Metall.*, 1953, 1:664-673.
8. Clapp PC, Moss SC. Correlation functions of disordered binary alloys. I. *Phys. Rev.* 1966, 142:418-427.
9. Khachaturian AG. *The theory of phase transformations and the structure of solid solutions.* Moscow, Nauka Publ., 1974, 256 c.
10. Lyakishev NP (Ed). *State diagrams of binary metal systems.* T. 3, Book 1. Mocsow, Mashinostroenie Publ., 1999, 872 c.

DOI: 10.17725/rensit.2020.12.241

Modified Sierpinski Carpet

Galina V. Arzamastseva, Mikhail G. Evtikhov, Feodor V. Lisovsky, Ekaterina G. Mansvetova

Kotelnikov Institute of Radioengineering and Electronics of RAS, Fryazino Branch, http://fire.relarn.ru/

Fryazino 141120, Moscow region, Russian Federation

E-mail: arzamastseva@mail.ru, emg20022002@mail.ru, lisf@df.ru, mansvetova_eg@mail.ru

Received March 25, 2020; reviewed April 6, 2020; accepted April 20, 2020

Abstract. **The algorithm is described and the properties of a previously unknown modification of the Sierpinsky carpet are studied. An example of proposed algorithm application for the fractal simulation of a really observed domain structure is given. An experimental study of light diffraction in the Fraunhofer zone was performed on computer-generated images of modified Sierpinsky carpets of different generations transferred to a transparent film using a high-resolution imagesetter with a small dot size. The observed diffraction patterns are compared with Fourier images of prefractals pictures approximated by the grid function.**

Keywords: **domain structure, diffraction pattern, Sierpinsky carpet, fractal, Fourier image, Hausdorf dimension**

UDC 51.74; 535.4

Acknowledgments. The work was carried out at the expense of budget financing within the framework of the state task.

For citation: Galina V. Arzamastseva, Mikhail G. Evtikhov, Feodor V. Lisovsky, Ekaterina G. Mansvetova. Modified Sierpinski Carpet. *RENSIT*, 2020, 12(2):241-246. DOI: 10.17725/rensit.2020.12.2.241.

Contents

1. INTRODUCTION

Previously, it was reported that a new fractal, which is a Sierpinsky carpet modification [1], was used to simulate a complex domain structure on the surface of uniaxial magnetic single crystalline plates [2]. In this paper, the properties of this fractal are considered in more detail, its Hausdorf dimension is determined, and the results of experiments on the observation of light diffraction by black-and-white bitmap images of different prefractal generations are presented.

The essence of fractal modification is illustrated in **Fig. 1**, where three successive stages of construction of the classic (upper row) and modified (lower row) Sierpinsky carpets are shown. The black color in the drawings is used for displaying fractal elements, and the white color

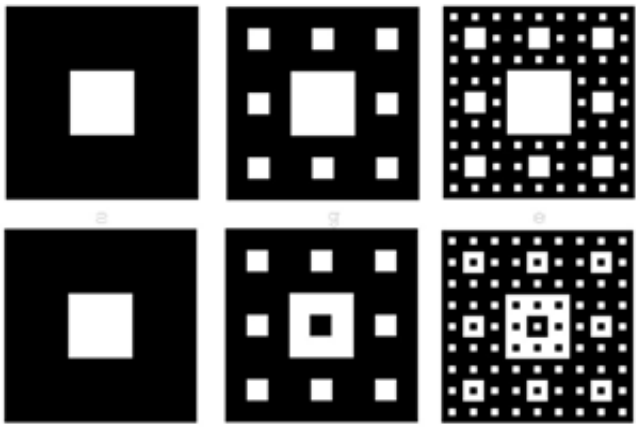


Fig. 1. *Three successive construction stages of the classic (upper row) and modified (lower row) Sierpinsky carpets.*

is used for displaying voids (holes). For a classic carpet, in the first step, the black square with sides of l is divided into nine equal-sized squares with sides of $l/3$, and the middle square is thrown out; then a similar procedure is performed with each of eight black squares bordering the central "white" square, which remains unchanged. Then the process is repeated for sixty-four squares with sides $l/3^2$, etc.

The construction of the modified Sierpinsky carpet also begins with a solitary black square, from which the central square with sides of $l/3$ is removed, but in the second step, in contrast to the classic carpet, a black square with sides of $l/3^2$ is inserted into the center of the figure. The third step is to divide the square into 81 squares in increments $l/3^2$ and then place a square with sides of $l/3^3$ the opposite color (black in white and vice versa) in the center of each of them. This process can be continued indefinitely.

In contrast to the classic Serpinsky carpet, which is a connected topological set, the modified carpet does not have connectivity. Another distinctive feature of this carpet is the lack of strict self-similarity, since the central square for any generation of prefractal differs from all others by inversion of color.

The Hausdorff dimension of a modified Sierpinsky carpet is defined by a formula $D_f = \lim_{n\to\infty} d_n$, where $dn = lnN_n/nln3$, and N_n is the number of black squares with a length of $l/3^n$ in the prefractal of the n-th order. To simplify the calculation (assuming that the value of l is equal to one without limiting generality) let's assume that the central square with a linear size 1/3, which is an inverse carpet of the n-th order in relation to the original, moves and overlays any of the eight side squares with the same linear size 1/3. The result is an entirely black square, where the number of black squares with a linear size $1/3^n$ is equal to $3^{2(n-1)}$. Each of the other seven squares with a linear size of 1/3 is a modified cover (n-1)-th order, that is, the number of black squares with a linear size $1/3^n$ in each of them is equal to N_{n-1}. There is a recurrent relation $N_n = 3^{2(n-1)} + 7N_{n-1}$, which implies that

$$N_n = 3^{2(n-1)} + 7\cdot 3^{2(n-2)} + 7\cdot 3^{2(n-3)} + \ldots +$$
$$+7^{n-2}\cdot 3^2 + 7^{n-1}\cdot 8 =$$
$$= 3^{2(n-1)}\cdot\left[1+\frac{7}{9}+\left(\frac{7}{9}\right)^2+\ldots\left(\frac{7}{9}\right)^{n-2}+8\left(\frac{7}{9}\right)^{n-1}\right] = \quad (1)$$
$$= 3^{2(n-1)}\left[\sum_{k=0}^{k=n}\left(\frac{7}{9}\right)^k + 8\left(\frac{7}{9}\right)^n\right].$$

Using the dependence $N_n(n)$ we find that the dimension of the modified Sierpinsky carpet is $D_f = \lim_{n\to\infty}\frac{2(n-2)\ln 3+\ln\left[9/2+8(7/9)^n\right]}{n\ln 3} = 2,$ that is, as the order of the prefractal increases, the entire source square is filled in, in contrast to the classic Sierpinsky carpet whose dimension is $ln8/ln3 = 1.89$.

2. THE DIFFRACTION PATTERNS AND THE FOURIER-IMAGES OF THE MODIFIED SIERPINSKI CARPETS

An experimental study of the diffraction of a collimated light beam (with a wavelength of 0.63 microns) in the Fraunhofer zone was performed on computer-generated black-and-white raster images of modified Sierpinski carpets of various generations transferred to a transparent film using a imagesetter with a resolution of 1333 points per centimeter (3386 dpi) and a point size of 7.5 microns. The image of the diffraction pattern on the screen in the diffraction plane was recorded using a digital camera. More detailed methods and features of the described experiments are described in [3].

For numerical determination of Fourier images, black-and-white bitmap images of modified Sierpinsky carpets were approximated by a grid function on a quadrate grid with a number of nodes $n_1 \times n_2$, where the values n_1 и n_2 were chosen sufficiently large (up to 4096) to adequately approximate the smallest-size prefractal details (in computer representation) and to enable the study of prefractals with high generation numbers. In our experiments, specific values were chosen so that the parameter p, that is equal to the ratio of the overall linear size of the smallest element of the prefractal to the grid period, was at least 9. For the image digitized in this way, the fast Fourier transform was used to determine the values of the quadrate of the Fourier component modules, i.e., the spectral distribution of the intensity of diffracted radiation in the Fraunhofer zone. To display the intensity of diffraction maxima on the plane, circles with a radius proportional to the intensity (or intensity logarithm) were used [3].

It was found that for modified Sierpinsky carpets, there is a marked difference between their Fourier images, i.e., diffraction patterns calculated from fractal pictures, and those observed in experiments. The central (fractal) parts for all diffraction

patterns are almost identical, but the peripheral ("lattice") parts are slightly different. In the experimental picture, there are reflexes from a certain square lattice, which are weakly reflected on the calculated diffractograms. It was found that this discrepancy is due to the difference in the size of isomorphic black and white squares, which was confirmed by observing a model with the image of a fractal on a transparent film under a microscope. When printing with a imagesetter, the laser beam partially illuminates the area outside of the formed ("black") image, as a result of which the black areas (squares) increase in size, and the white (not illuminated) areas, on the contrary, shrink. This difference in size is minimal for large white and black squares and maximal for the smallest squares, where diffraction mainly forms the peripheral part of the diffraction pattern.

The described difference between Fourier images and the experimental diffraction patterns for a modified Sierpinsky carpet depends on the coefficient $c_{w/b} = r_w/r_b$, where r_w and r_b, are the linear dimensions of the smallest white and black squares on a transparent film respectively. This is illustrated in **Fig. 2**, which shows experimentally obtained (left) and calculated diffraction patterns at $c_{w/b}$ = 1.0 (center) and $c_{w/b}$ = 0.64 (right) for a modified carpet of the 6th order. This can be used to reduce the distortion described above by artificially reducing the size of the black squares on the prefractal bitmap images. For the example in Fig. 2 on the right is a calculated diffractogram for a modified Sierpinsky carpet of the 6th order at $c_{w/b}$ = 0.64, which corresponds well to the experimental shown on the left, in contrast to the diffractogram in the center for the value $c_{w/b}$ = 1.0.

The classic Sierpinski carpet is formed by sublattices consisting only of black squares, so, despite the above-mentioned system error of the imagesetter, the experimental one (on the left in **Fig. 3**) and calculated (on the right in Fig. 3) diffractograms correspond well to each other.

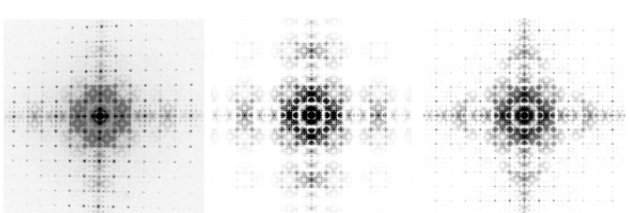


Fig. 2. *Experimentally obtained diffraction pattern (left) and calculated diffraction patterns for (center) and for (right) for a modified Sierpinsky carpet of the 6th order.*

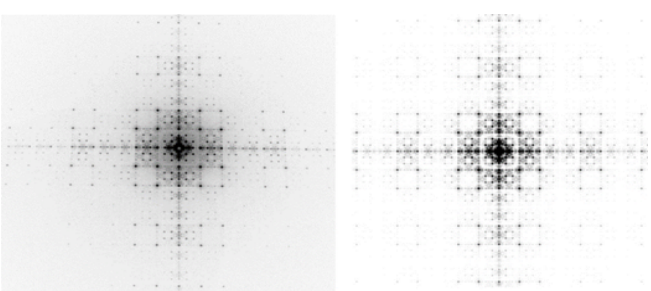


Fig. 3. *Experimentally obtained diffraction pattern (left) and calculated diffraction patterns (right) for a classic Sierpinsky carpet of the 6th order.*

The central part of the diffraction pattern for the modified Sierpinsky carpet has spatial invariance (with a scaling factor equal to 3), as well as for the classical carpet.

Analysis of the diffraction pattern allows us to find the Hausdorff dimension D_f using the circle method (see for example [2]), based on the numerical determination of the average resulting intensity of diffracted radiation $\bar{I}$ in circles with a center at the location of the main diffraction maximum and with a variable radius equal to $r_k = r_0 + k\delta_r$, where r_0 and δ_r is the initial radius and the step of radius change, $k = 0,1,2,...$ and sequential use of the next formula $\bar{I}(r_k) = I_0 \exp(-D_f)$, from which it follows that the fractal dimension is equal to the modulus of the angular coefficient of the line approximating the dependence $\bar{I}(r_k)$ on the double logarithmic scale.

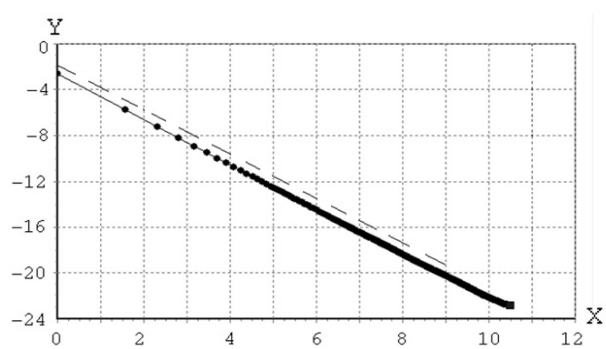


Fig 4. *Dependence for a modified Sierpinsky carpet of the 8th order with the use of normalized variables* $Y = \ln \bar{I}/\ln 2$ *and* $X = lnr_k/ln2$ *on a double logarhyme scale (angular coefficient module of the dashed line equals 1.935).*

For the modified Sierpinsky carpet of the 8th order, the dependence $\bar{I}(r_k)$ for which is on double logarithmic scale using normalized variables $Y = \ln \bar{I}/\ln 2$ and $X = lnr_k/ln2$ is shown in Fig. 4, the value of D_f was 1.936, which is significantly different from 2. The reason is that the value of d_n converges very slowly to the value 2. So, for $n = 8$ the value of D_f equals 1.935, that well corresponds to a certain dimension value (see the dashed line in Fig.4 with module angular coefficient equal to 1.935). Note that the coating method (see e.g. [4]) for this prefractal gives value $D_f = 2$.

3. CONCLUSION

The main results of the work performed are summarized as follows. For the geometric description of a complex domain structure on the surface of the uniaxial magnetic singlecrystal plates, an algorithm based on the use of a previously unknown modification of the Sierpinsky carpet was developed. An example of application of the proposed algorithm for simulation a real observed domain structure is given. An experimental study of light diffraction in the Fraunhofer zone was performed on computer generated images of modified Sierpinsky carpets of different generations transferred to a transparent film using a high-resolution imagesetter with a small dot size. The observed diffraction patterns

were compared with Fourier images of prefractals pictures approximated by the grid function. The difference between experimental and calculated diffraction patterns for the modified Sierpinsky carpet was found to be an artifact, and the reason for this difference was revealed.

The work was carried out at the expense of budget financing within the framework of the state assignment.

REFERENCES

1. Sierpiński W. Sur une courbe cantorienne qui contient une image biunivoque et continue de toute courbe donnée. *C. R. Acad. Sci. Paris*, 1916, 162(25):629–632.
2. Arzamastseva G.V., Evtikhov M.G., Lisovskii F.V., Mansvetova E.G. Fraktalnaya model slozhnoy pripoverkhnostnoy domennoy struktury vysokoanizotropnykh odnoosnykh monokristallov [Fractal model of a complex near-surface domain structure of highly anisotropic uniaxial single crystals]. *Physics of Metals and Metallography,* 2020, 121(5):454-457 (in Russ.).
3. Arzamastseva G.V., Evtikhov M.G., Lisovskii F.V., Mansvetova E.G. *RENSIT*, 2017, 9(2):221-229. DOI: 10.17725/rensit.2017.09.221.
4. Feder J. *Fractals*. New York, 1986, 283 p.

DOI: 10.17725/rensit.2020.12.247

Physical and electrodynamic properties of nanoscale conductive films on polymer substrates

Alim S.-A. Mazinov

Vernadsky Crimean Federal University, https://cfuv.ru/

Simferopol 295007, Russian Federation

E-mail: mazinovas@cfuv.ru

Received March 27, 2020; reviewed April 6, 2020; accepted April 20, 2020

Abstract. **The interaction of high-frequency electromagnetic radiation with thin nanometer-sized aluminum films deposited by magnetron sputtering on flexible polymer substrates has been considered. Experimental studies were carried out on a vector panoramic meter P4226 in the frequency range of 8.2 - 12.2 GHz. The paper reveals the dependences of the relative powers of the reflected, absorbed, and transmitted waves on the film thickness and the relationship between the electrodynamic characteristics and the surface topography.**

Keywords: **microwave radiation, nanoscale films, metal-dielectric structures, absorption coefficient, reflection, transmission, aluminum**

PACS 61.48.+c, 61.66.Hq, 73.61.−r

For citation: Alim S.-A. Mazinov. Physical and electrodynamic properties of nanoscale conductive films on polymer substrates. *RENSIT*, 2020, 12(2):247-252. DOI: 10.17725/rensit.2020.12.247.

Contents

1. INTRODUCTION

The steady increase in the number of mobile radio-electronic equipment: radio stations, smartphones and many other devices [1], together with an increase in intelligence capabilities and a significant reduction in power consumption, results from a decrease in the size of the elements on the integrated circuit (IC), in which conductive tracks and pads make up a significant part [2]. At present, industrial technologies for the production of ICs involve the use of active devices, the sizes of which reach tens or several nanometers [3]. Therefore, the use of conductive films of nanometer thicknesses in the production of electronic components is a future-oriented trend, though further studies of their electrodynamic properties are still needed.

On the other hand, studying the electrodynamic properties of the films are interesting in view of their inclusion in complex metostructures [4], which make it possible to achieve absorption of 80–90% [5]. It should be noted that pure conductive films with thicknesses from 2 to 7 nm can significantly (up to 50%) convert the electromagnetic fields energy of the radio range of 1 ... 400 GHz [6,7] into thermal energy. Although this property of the films limits their use in "conductive" electronics, such resistive properties are useful for stealth technologies [8], as well as for creating thin-film filters, protective shields, or specialized sensors [9].

The topology of the dielectric-metal interface is one of the key issues in creating such absorbing structures as it determines the specifics of the electrodynamic properties of the structure as a whole [10]. Most of the research work on this topic deals with the

study of the interaction of radiation and metal-dielectric structures based on solid substrates [11, 12]. However, the prospects for creating flexible devices, as well as the possibility of simplifying the production of flexible screens and antennas, require special consideration of the interaction of electromagnetic radiation with a metal-dielectric structure on polymer substrates [13]. Consequently, the aim of the present paper is to identify the specifics of the interaction between electromagnetic radiation from the microwave range and the MDS, in relation to the surface geometry of the flexible substrate.

2. THE EXPERIMENTAL RESEARCH TECHNIQUE AND THE OBJECT OF THE RESEARCH

Studies of the relative powers of reflected, transmitted, and absorbed waves (optical coefficients) in the range 8.2–12.2 GHz were carried out using the Mikran P4226 vector network analyzer (**Fig. 1**). To obtain normalized data and calculate the losses of the waveguide system, at the beginning and end of the experiment, the measurements of the parameters on clean flexible substrates were made with an attached 200 μm thick aluminum plate, exceeding the skin layer for the conductor at the lower boundary of the working frequency range. To compensate for the influence of coaxial-waveguide transitions, as well as other factors, calibration was performed using a reflection measure and a quarter-wave line (TRL: Thru, Reflect, Line) [14], which made it possible to obtain fairly accurate results.

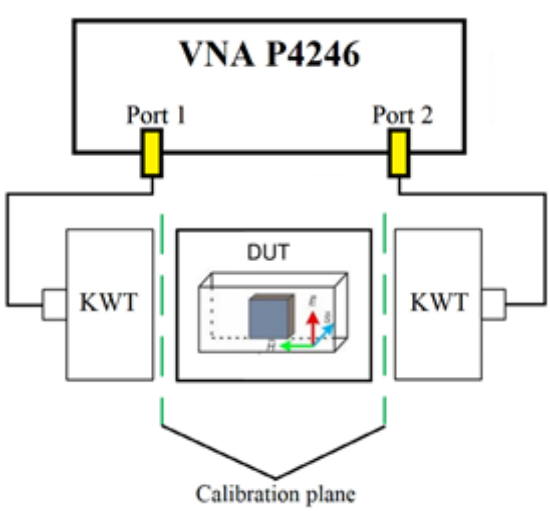


Fig. 1. *Installation block diagram.*

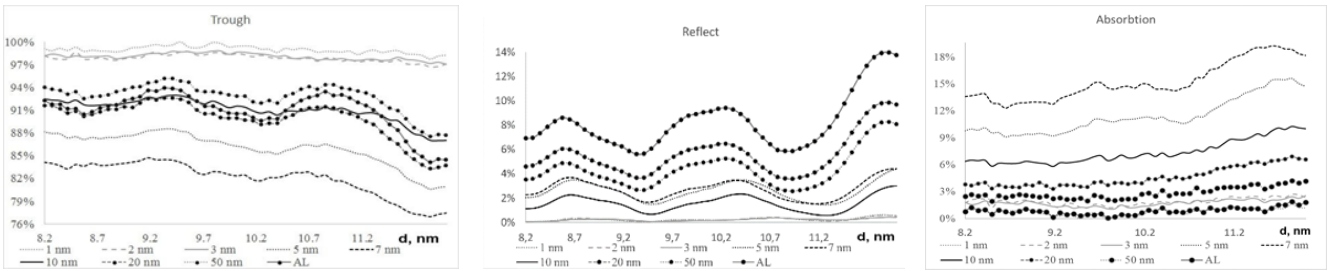
The direct interaction between the microwave and the samples was determined by the matrix of S-parameters, the main components – S21 and S11 – corresponding to a direct drop from the first port, were chosen. As the initial measurements showed, the properties of the measuring waveguide path, with the structure under study, are close to the properties of a reversible four-terminal circuit, i.e. the transmission coefficient is the same in both directions. Accordingly, we used the main components S21 and S11, corresponding to a direct drop from the first port of the vector analyzer. The coefficients of transmitted, reflected and absorbed power were determined using the obtained S-parameters (Fig. 2):

$$T = \frac{P_{thru}}{P_{inc}} = \frac{|V_{thru}|^2}{|V_{inc}|^2} = |S_{21}|^2,$$

$$R = \frac{P_{reflect}}{P_{inc}} = \frac{|V_{reflect}|^2}{|V_{inc}|^2} = |S_{11}|^2,$$

$$A = 1 - |S_{11}|^2 - |S_{21}|^2.$$

Metal-dielectric structures were fixed in the geometric center of the waveguide path, perpendicular to the axis of the waveguide with standard dimensions of 23×10 mm. To avoid the capacitive and inductive effect of metallization on the measuring system, the effective area of interaction of radiation and the samples was 10% of the waveguide area - 6×6 mm. The samples were placed on the geometric center of the waveguide cross section and were fixed by a dielectric substrate made of synthetic foam, "transparent" for electromagnetic radiation. Thus, the sample was affected by the maximum power of H10



a b c

Fig. 2. *Frequency dependences of the coefficients: a - passage; b - reflection; c - absorption.*

wave electric field and the major part of the wave energy interacted with a nanoscale conducting film.

Films made of lavsan (polyethylene terephthalate), fluoroplastic and teflon were taken as the basis for MDS. This paper presents the results of studying MDS with lavsan films, since the results involving the study of fluoroplastic and teflon substrates differ insignificantly. The conductive part was deposited by the method of magnetron sputtering of the target using the URM 3.1279.0129 installation with additional ion beam processing. What is more, the planetary rotation of the substrates ensured uniform coating deposition, and in combination with adjusting the deposition time, it was made possible to obtain a range of the conducting layers thicknesses: 1, 3, 5, 7, 10, 20, 50 nm.

The initial assessment of the substrates surfaces of was carried out using an optical interference microscope LOMO MII-4M in broadband radiation of a white LED with subsequent detailing of the relief using red and blue lasers. A more detailed analysis of the formation process and morphological characteristics of aluminum nanofilms was carried out using an NT-MDT nanoeducator 2 scanning atomic force microscope, to compare them with the absorption, transmission, and reflection spectra.

3. DISCUSSION OF THE RESULTS

Along with a number of advantages, there is one more benefit to the method of using 10% of the interaction area in the waveguide – its significant sensitivity to diffraction. As a result, a monotonic decrease in the transmission coefficient T (Fig. 2*a*) and an increase in the reflection coefficient R (Fig. 2*b*) for thin conducting films do not correlate with theoretical calculations, which show independence of T, R on the frequency and, accordingly, independence of the absorption A on the frequency as well [6,15]. This is not due to the specifics of the physical properties of active MDS, but is accounted for by a decrease in the diffraction properties of the object under study with a decrease in the wavelength.

In accordance with the established concepts, the reflection coefficient should vary from the minimum value corresponding to the substrate without metal ($d = 0$), and monotonically increase to the value corresponding to the formed film ($d > 15 \ldots 20$ nm). However, in the range of film thicknesses $5 < d < 10$ nm, there is a deviation from the monotonicity of the reflection coefficient growth (Fig. 2*b*), which is associated with the relief of the film.

The maximum absorption of an aluminum film on lavsan is achieved at the conducting layer thickness $d = 7$ nm (Fig. 2*c*). This is due to the transition of the MDS from the dielectric ($d = 0$) with the geometric dimensions of 6×6 mm, to film structures of greater thickness, causing electric short circuiting of the space. This leads to a change in conductivity from 0 to values characteristic of the bulk of a continuous material, in our case Al $\sim 3.8 \cdot 10^7$ S/m. It is with the thickness $d = 7$ nm that the specific conductivity reaches about 10^6 S/m [6,

9], and there occurs the maximum conversion of the induced currents to thermal energy.

For an in-depth understanding of the nature of the nonlinear interaction effect at the thickness of 7 nm, the growth dynamics of a conductive material on a flexible substrate was analyzed.

4. THE CORRELATION BETWEEN ELECTRODYNAMIC CHARACTERISTICS AND FILM MORPHOLOGY

A general analysis of surface morphology, carried out by a sequential study of optical microscopy with subsequent analysis of AFM images, revealed a more complex surface morphology of MDS on flexible substrates compared to solid ones [16]. From an aggregate analysis of profilograms, it follows that the polymeric substrate made of lavsan has relatively large height differences (**Fig. 3***a*), however, in contrast with solid amorphous substrates, the change in height occurs rather slowly and smoothly, within small areas, and the surface is quite smooth (PTFE and Teflon films have similar the reliefs). It should be noted that the surface of the substrate was not pre-processed.

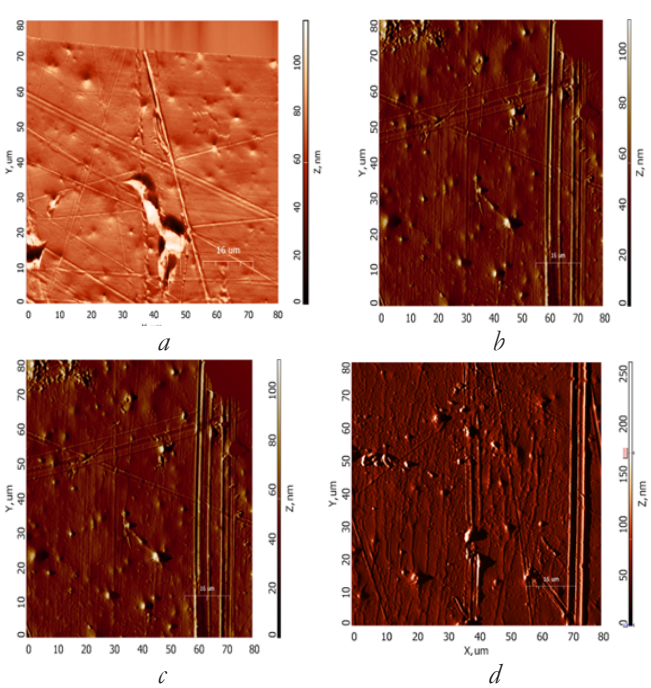


Fig. 3. *Dynamics of changes in the surface morphology using a border detection filter (Pruitt filter): a - clean substrate; b - 5 nm aluminum film; c - thickness 7 nm; d - thickness of 10 nm.*

After deposition of aluminum with the thickness of 5 nm (Fig. 3*b*), lateral micro-formations, reaching tens of nanometers in height, are visible on the substrate. With longer spraying, the number of conductive micro-formations increases significantly (Fig. 3*c,d*). With an approximate coating thickness of 10 nm (Fig. 3*d*), the formation of a smoother surface in observed in the AFM image, with elevation drops larger than 7 nm.

The quantitative analysis of the correlation between the electrodynamic parameters of the radiation effects and the surface topology was based on statistical data and the root-meansquare value of roughness (*Zq*), which was determined by measuring the value of the deviations from the midline:

$$Zq = \left(\frac{1}{N} \sum_{j=1}^{N} z_j^2 \right)^{1/2},$$

where z_j is the deviation of the j-th point, N is the number of points in the image.

A non-standard increase in the absorption coefficient for films with thicknesses of 7 nanometers correlates quite well with the dependence of *Zq* on their thickness (**Fig. 4**). A sharp increase in the root-meansquare

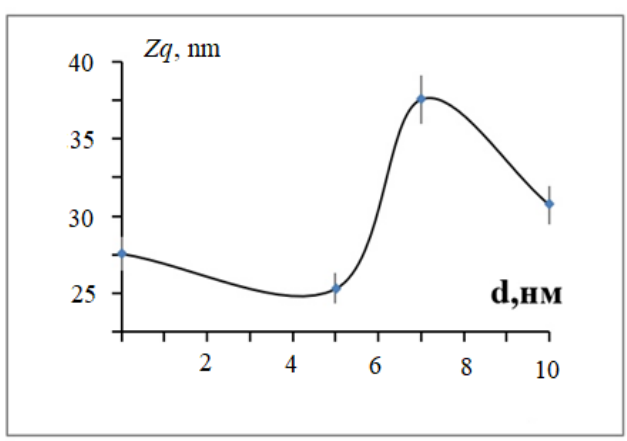


Fig. 4. *The dependence of the change in the root mean square roughness on the thickness of the coating.*

value of roughness of the coating indicates a transient process of the conversion of localized nano and micro objects into a continuous conductive layer, which leads to an increase in the total conductivity. An intermediate step in this process is the formation of "bridges" between metal formations, which is the reason for the growth of *Zq*. However, the average conductivity of the floor bridges at the initial stage of growth is small, as is the conductivity of the crystallites themselves at small sizes, taking into account possible oxidative processes. Therefore, part of the induced incident microwave energy of alternating current is transformed into Joule heat on a non-ideal conductor.

The increase in the reflected wave energy at thicknesses of more than 10 nm (Fig. 2*b*) is accounted for by the increase in induced currents, which include inductive and capacitive components. The general conduction currents, in turn, are due to the emergence of stable galvanic bonds between individual conductive islands and the primary growth of connecting "bridges", as confirmed by a sharp decrease in *Zq* at an increase in the conductive layer thickness (Fig. 4). A further process of the conductive material deposition leads to the formation of a continuous conductive film having not only a uniform atomic structure, but also sufficient uniformity in thickness, which is exactly due to the relatively smooth surface of the substrate. This, in turn, leads to a sharp increase in conductivity, approaching the conductivity of a bulk material, which leads to an increase in the conversion of the incident wave into the reflected wave by means of increased currents.

5. CONCLUSION

The study of the interaction of EMR and MDS on flexible substrates showed the identical nature of absorption, as well as it did in their solid-state counterparts, the main feature being the absorption coefficient peak occurring within a small range of the conductive layer thickness. A feature of the interaction of electromagnetic radiation and a flexible layered structure of aluminum-lavsan is the resonance at the thickness of 7 nm, which is greater than that of similar solid structures.

In the case of our study, the key point determining the reflection and absorption of MDS was the modification of the resistive properties of the aluminum conductive layer. It is the specifics of the lavsan surface geometry, where smoother peaks were observed between the heights of 100-150 nm, that formed the conductive layer by means of islands, which were connected by bridges as they increased. Exactly at the instance of the conductive bridges formation in the structure, the conductivity arises, converting the energy of the incident radiation into joule heat. Further growth of the conductive layer leads to smoothing of the relief and forming a continuous structure with conductivity approaching that of the bulk, thereby increasing the reflected wave fraction.

REFERENCES

1. Konstantinou D, Bressner TAH, Rommel S, Johannsen U, Johansson MN, Ivashina MV, Smolders AB, Monroy IT. 5G RAN architecture based on analog radio-over-fiber fronthaul over UDWDM-PON and phased array fed reflector antennas. *Optics Communications,* 2020, 454:124464, doi: 10.1016/j.optcom.2019.124464.
2. Hills G, Lau C, Wright A. Modern microprocessor built from complementary carbon nanotube transistors. *Nature,* 2019, 572:595-602, doi: 10.1038/s41586-019-1493-8.
3. Khan HN, Hounshell DA, Fuchs ERH. Science and research policy at the end of Moore's law. *Nature Electronics,* 2018, 1(2):146–146, doi: 10.1038/s41928-017-0005-9.

4. Yuan J, Liu Q, Li S, Lu Y, Jin S, Li K. Metal organic framework (MOF)-derived carbonaceous Co 3 O 4 /Co microframes anchored on RGO with enhanced electromagnetic wave absorption performances. *Synthetic Metals,* 2017, 228:32-40, doi: 10.1016/j.synthmet.2017.03.020.
5. Tran MC, Pham VH, Ho TH, Nguyen TT, Do HT. Broadband microwave coding metamaterial absorbers. *Scientific Reports,* 2020, 10(1):1810, doi: 10.1038/s41598-020-58774-1.
6. Nimtz G, Panten U. Broad band electromagnetic wave absorbers designed with nano-metal films. Ann. Phys. (Berlin), 2010, 19(1-2): 53-59. DOI 10.1002/andp.200910389
7. Li S, Anwar S, Lu W, Hang ZH, Hou B, Shen M, Wang CH. Microwave absorptions of ultrathin conductive films and designs of frequency-independent ultrathin absorbers. *AIP Advances*, 2014, 4(1):017130, doi: 10.1063/1.4863921.
8. Ahmad H, Tariq A, Shehzad A, Faheem MS, Shafiq M, Rashid IA. Stealth technology: Methods and composite materials-A review. *Polymer Composites,* 2019, 1:16. DOI: 10.1002/pc.25311.
9. Vdovin VA. Nanometer metal films in the sensors of powerful microwave pulses. *III All-Russian Conference "Radar and Radio Communication". IRE RAS*, 2009, 832-835.
10. Fitaev IS, Orlenson VB, Romanets YV, Mazinov AS. Surface topologies of thin aluminum films and absorbing properties of metal dielectric structures in the microwave range. *ITM Web of Conferences,* 2019, 30:08013, doi: 10.1051/itmconf/20193008013.
11. Andreev VG, Vdovin VA, Pronin SM, Khorin IA. Measurement of optical coefficients of nanometer metal films at a frequency 10 GHz. *Journal of Radio Electronics (IRE RAS),* 2017, 11:4.
12. Zuev SA, et al. Microwave Range Diffraction Properties of Structures with Nanometer Conductive Films on Amorphous Dielectric Substrates. *26th Telecommunications Forum (TELFOR*), 2018, 1-4, doi: 10.1109/TELFOR.2018.8611867.
13. Antonets IV, Kotov LN, Makarov PA, Golubev EA. Nanostructure, conductive and reflective properties of thin films of iron and $(Fe)_X(BaF_2)_Y$. *Journal of Technical Physics*, 2010, 80(9):134-140.
14. Guba VG, Ladur AA, Savin AA. Classification and analysis of calibration methods for vector network analyzers. *TUSUR Reports*, 2011, 2(24):149-155.
15. Orlenson VB, Zuev SA, Starostenko VV. *ITM Web of Conferences CriMiCo'2019,* 2019, 30:08011, doi: 10.1051/itmconf /201930 ITM.
16. Starostenko VV, Mazinov AS, Fitaev IS, Taran EV, Orlenson WB. The dynamics of the surface formation of conductive aluminum films on amorphous substrates. *Applied Physics*, 2019, 4:60-65.

DOI: 10.17725/rensit.2020.12.253

Inverse problems of macrofracture formations exploration seismology solution with use of convolutional neural networks

Maxim V. Muratov, Vasily V. Ryazanov, Igor B. Petrov

Moscow Institute of Physics and Technology, https://mipt.ru
Dolgoprudnyi 141700, Moscow region, Russian Federation

E-mail: max.muratov@gmail.com, vassavadda@gmail.com, petrov@mipt.ru

Received July 3, 2020, peer reviewed July 07, 2020, accepted July 08, 2020

Abstract. **This article is devoted to solving the inverse problems of exploration seismology of uniformly oriented macrofractures systems using convolutional neural networks. The use of convolutional neural networks is optimal due to the multidimensionality of the studied data object. A training sample was formed using mathematical modeling. In the numerical solution of direct problems, a grid-characteristic method with interpolation on unstructured triangular meshes was used to form a training sample. The grid-characteristic method most accurately describes the dynamic processes in exploration seismology problems, since it takes into account the nature of wave phenomena. The approach used makes it possible to construct correct computational algorithms at the boundaries and contact boundaries of the integrational domain. Fractures were set discretely in the integration domain in the form of boundaries and contact boundaries. The article presents the results of solving inverse problems with variations in the angle of inclination of fractures, height of fractures, density of fractures in the system, as well as joint variations in the angle of inclination and height of fractures and all three investigated parameters.**

Keywords: **machine learning, convolutional neural networks, mathematical modeling, grid-characteristic method, exploration seismology, inverse problems, fractured media**

UDC 004.93

Acknowledgments. This work was carried out as part of the RSF project No. 19-11-00023 based on MIPT.
For citation: Maxim V. Muratov, Vasily V. Ryazanov, Igor B. Petrov. Inverse problems of macrofracture formations exploration seismology solution with use of convolutional neural networks. *RENSIT*, 2020, 12(2):253-262. DOI: 10.17725/rensit.2020.12.253.

Contents

1. INTRODUCTION

Currently, exploration seismology [1,2] is one of the most reliable methods for finding oil and gas deposits and preparing the rock before deep drilling. The ongoing research is aimed at determining the structure of geological

formations, as well as the possible location of the hydrocarbon field. Dense carbonate rocks and deep-lying sandstones account for an increasing share of exploration. Hydrocarbon-containing formations of such rocks are usually penetrated by systems of subvertical, uniformly oriented fractures of various scales, filled with liquid [3,4]. They determine the reservoir properties, being the basis for constructing models of fields to justify their development modes.

One of the founders of the theory of seismic migration was J. F. Claerbout [5,6]. With the advent of modern high-performance computing systems, considerable efforts have been made to develop new high-precision methods [7,8] for solving inverse seismic problems. Initially, all methods were based on an acoustic approach that did not take into account the influence of shear waves. To overcome this drawback, a two-wave elastic model was used [9]. Nowadays the great interest for seismologists is the identification of fractured zones. This is due to the high permeability of the medium and the potentially high content of hydrocarbons. Various mathematical models have been developed taking into account the complex structure of the geological medium [10-12]. Diffracted waves are in the process of a comprehensive study by exploration geophysicists. A large number of scientific papers are devoted to numerical modeling of seismic reactions from fractured media [13,14].

In recent years, machine learning methods, and in particular deep neural networks, have shown impressive results in many areas, such as computer vision, speech recognition, and machine translation. For example, in the field of computer vision, it was possible to solve many problems that were previously unsolved, such as the classification problem [15], the recognition problem [16], and the image generation problem [17].

One of the significant advantages of deep learning methods is that these methods can be transferred to many other areas related to the processing of large amounts of data. One such area is the task of seismic exploration. Several works in this area have already been carried out. In [18], the problem of detecting a fault in 2 dimensions using a deep convolutional neural network was solved. Synthetic data obtained by solving large direct problems were used as data for training the neural network. In [19], a similar problem was solved in 3 dimensions. The great advantage that is highlighted in these works is that the input data for deep learning algorithms do not require special processing and, therefore, such methods can be easier to use than standard seismic methods. Flexibility and relative simplicity make such methods effective to solve practical problems. So, in [20], deep neural networks are used to detect CO_2 emissions, and in [21], these methods are used to detect and classify defects in composite materials.

Formulation of the problem

In this paper, we consider the process of solving the inverse problem of exploration seismology of systems of unidirectional macrofractures (**Fig. 1**) using convolutional neural networks. The following notation is introduced in the figure: α – fracture inclination angle, h – the height of fractures, d – the distance between fractures (a parameter characterizing the density of fractures at a given horizontal extent of the system). The horizontal dimension in all experiments was considered constant and equal to 1000 m. The angle of inclination of

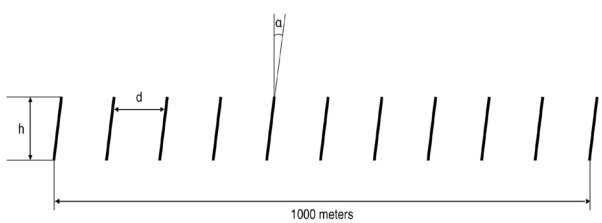


Fig. 1. *The scheme of macrofractures system.*

the fractures varied in the interval from -15° to +15° relative to vertical, the height of fractures varied from 50 to 200 m, the distance between fractures – from 50 to 200 m.

The training set was formed using mathematical modeling of direct two-dimensional problems using the grid-characteristic method on unstructured triangular meshes [22].

In the considered inverse problems, the elastic characteristics of the geological medium were considered known: longitudinal wave velocity 2250 m/s, transverse – 1250 m/s, the density of medium – 1180 kg/m^3. The system is placed in depth of 2000 m. A plane wave was excited on the surface of the earth, propagating deep into the geological medium. The response of the reflected and diffracted waves, which were formed when the incident wave front passed through the system of fractures, was recorded on seismic receivers located on the surface above the system of fractures with an interval of 50 m (100 receivers in total).

The problems in which the following parameters were varied were considered:

- the angle of inclination of the fractures;
- the height of the fractures;
- the angle of inclination and the height of the fractures at the same time;
- fracture density (distance between fractures);
- simultaneously, all three parameters - the angle, height and density of fractures.

Qualitatively, the influence of the height and distance between fractures on the response picture was described in [23].

2. MATERIALS AND METHODS

2.1. Methodology for direct problem solving

The determining equations system of a linearly elastic medium can be represented in the form [24]:

$$\rho\frac{\partial V_i}{\partial t}=\frac{\partial T_{ji}}{\partial x_j},\ \frac{\partial \mathrm{T_{ij}}}{\partial \mathrm{t}}=\lambda\left(\sum_k\frac{\partial V_k}{\partial x_k}\right)I_{ij}+\mu\left(\frac{\partial V_i}{\partial x_j}+\frac{\partial V_j}{\partial x_i}\right). \quad (1)$$

where V_i – the velocity component, T_{ji} – the stress tensor, ρ – the density of medium, λ and μ – the Lame coefficients, I_{ij} – the component of unit tensor. By entering the vector of variables $\vec{u}=\{V_x,V_y,T_{xx},T_{yy},T_{xy}\}$, the system (1) write as:

$$\frac{\partial\vec{u}}{\partial t}+\sum_{i=1,2}A_i\frac{\partial\vec{u}}{\partial\xi_i}=0. \quad (2)$$

The numerical solution of (2) is found using the grid-characteristic method [25]. We carry out coordinatewise splitting and by changing variables we reduce the system to a system of independent scalar transfer equations in Riemann invariants:

$$\frac{\partial\vec{w}}{\partial t}+\Omega_i\frac{\partial\vec{w}}{\partial\xi_i'}=0,\ \ i=1,2. \quad (3)$$

For each transfer equation (3), all nodes of the computational mesh are bypassed, and characteristics are omitted for each node. From the time layer n, the corresponding component of the vector $\vec{w}$ transfer to the time layer n+1 as

$$w_k^{n+1}(\xi'_i)=w_k^n(\xi'_i-\omega_k\tau),$$

where τ – the time step.

After all the values are transferred, we go to a reverse transition to the vector of the desired values $\vec{u}$.

The interpolation on unstructured triangular meshes is considered. Values at each point are found using values at mesh reference points $\vec{w}(\vec{r}_{ijkl})$ and their weights $p_{ijkl}(\vec{r})$ as:

$$\vec{w}(\vec{r})=\sum_{i,j,k,l}\vec{p}_{ijkl}(\vec{r})\,\vec{w}(\vec{r}_{ijkl}).$$

The grid-characteristic method allows the most correct algorithms to be applied at the boundaries and contact boundaries of the integration domain [26,27].

The boundary condition can be written in common view as:

$$\boldsymbol{D}\vec{u}(\xi_1,\xi_2,t+\tau)=\vec{d},$$

where $\mathbf{D}$ – some matrix 5×2, $\vec{d}$ – some vector, $\vec{u}(\xi_1,\xi_2,t+\tau)$ – the values of the desired

velocity components and the components of the stress tensor at the boundary point at the next time step.

2.2. Mathematical models of fractures

In real exploration seismology problems, one has to deal with heterogeneity in the nature of the interaction of elastic waves with the surface of a fracture as it passes through it. A fracture is a complex heterogeneous structure [4,28]. In places, the flaps of the fractures are at some distance and are separated by a saturating fluid or gas [28], in some points adhesion is observed, the walls are closely adjacent to each other under the action of pressure forces [29]. In addition, fractures can be classified according to the nature of saturation: fluid or gas [28, 29].

In the problem under consideration, discrete fractures models based on the concept of an infinitely thin fracture were used. The fracture was defined as a boundary or contact boundary with a certain boundary condition.

a) Gas-saturated fracture

The gas-saturated fracture model simulates well the behavior of fractures filled with air or gas at a shallow depth of 100-150 m [29]. At great depths, under the influence of pressure, fractures with air close, and gas acquires the properties of a liquid.

The fracture is defined as the boundary condition of free reflection on the flap flaps:
$\boldsymbol{T}\vec{n} = 0.$

b) Fluid-filled fracture

In most practical problems, fractures are filled with fluid: water, oil, liquefied gas, etc. [22,28,29] Therefore, it was advisable to develop a model to describe such a situation.

A fluid-filled fracture is defined as a contact boundary with the condition of free sliding [22]:

$$\vec{v}_a \cdot \vec{n} = \vec{v}_b \cdot \vec{n}, \;\; \vec{f}_n^{\,a} = -\vec{f}_n^{\,b}, \;\; \vec{f}_\tau^{\,a} = \vec{f}_\tau^{\,b} = 0.$$

Such a contact boundary completely transmits longitudinal vibrations without reflection and completely reflects transverse waves. Such a picture corresponds to the real situation: the values of the propagation velocities of longitudinal waves in liquids and densities are comparable with the values of velocities and densities of geological media; while the rates of transverse vibrations in liquids are close to zero.

c) Glued fracture

At great depths under the influence of pressure, it happens that the flaps of the fractures touch so that the elastic waves almost completely pass through the fracture. In this case, it will optimally use the contact condition of complete adhesion [22]:

$$\vec{v}_a = \vec{v}_b, \;\; \vec{f}_a = -\vec{f}_b,$$

where $\vec{v}$ are velocities of closed boundary points, $\vec{f}$ – the force acting to the boundary, *a* and *b* are the first and the second flaps of fracture.

d) Partially-glued fracture

In real exploration seismology, partially glued fractures can occur [22,29], in which part of the surface of the flaps is sticky and part is separated by a fluid or gas. Such fractures show partial transmission of the elastic wave front, which affects the amplitudes of the response waves in the seismograms.

A model of the fracture was developed, where at different points of the flaps the conditions of gas-saturation (fluid-filling) and complete adhesion were randomly set. The number of certain points was regulated by a weight coefficient – the coefficient of gluing. Such a model made it possible to specify gas-saturated and fluid-filled fractures with a percentage of sticking points from 0 to 100% percent.

2.3. The neural network structure

In all experiments, a similar neural network architecture was used. Samples were generated

during the solution of the direct problem, the structure of the objects supplied to the input of the neural network coincided, the set of target variables differed.

Each object was a set of measurements of the horizontal and vertical components of the velocity (V_x, V_z) of vibrations. Velocity data was obtained from a series of evenly spaced sensors and measured at equal time intervals.

Based on this, each sample was transformed into a three-dimensional object (velocity component, number of sensors, number of time measurements) of size (2, 100, 300), respectively (**Fig. 2**).

To predict the angle of inclination, a convolutional neural network was used. Convolutional networks have worked well in solving problems of classification, regression, segmentation, etc. on visual-, audio- and other data.

The convolutional network differs from other types of neural networks by the presence of convolutional and pooling layers. These layers can significantly reduce the number of network parameters, accelerate the speed of learning.

The following types of layers were used in the current task: convolutional layer, MaxPooling layer, fully connected layer. Experiments were carried out with the addition of Dropout layers, but their use impaired the accuracy of

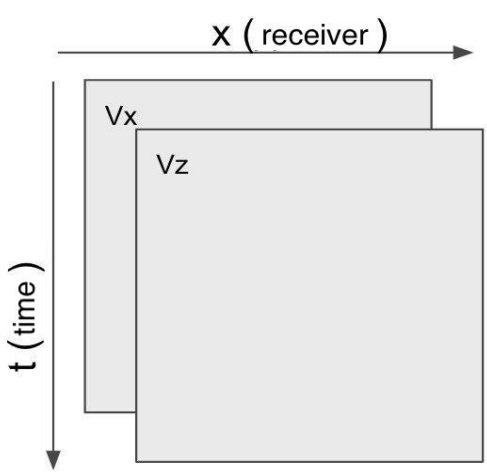


Fig. 2. *The scheme of learning and validation sets.*

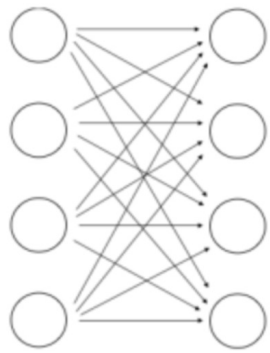


Fig. 3. *Fully connected layer.*

the predictions. Here is a brief description of each of the layers used.

a) Fully connected layer

The fully connected layer is a classic for most types of neural networks. In this layer, each neuron from the previous layer is connected to the neuron of the next layer (**Fig. 3**). Layers of this group are used in many types of tasks: their advantage is that they take into account the maximum amount of information and connections between neurons. The disadvantage is the large number of parameters, which is equal to the number of edges in conjunction with the number of output neurons. Another drawback is the fact that a large number of parameters can degrade the convergence of the optimized function.

b) Convolutional layer

Convolutional layers are a characteristic feature of the convolutional neural network (**Fig. 4**). Their feature is that instead of pairwise combining of outgoing and incoming neurons,

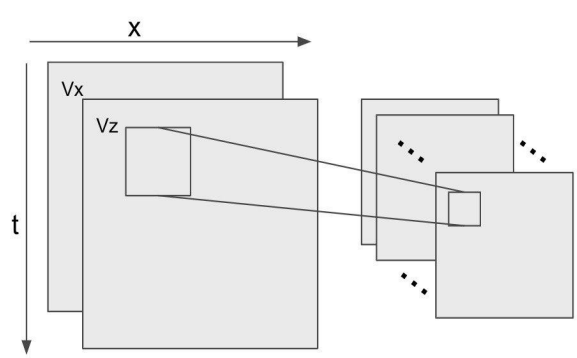


Fig. 4. *Convolutional layer.*

only certain local neurons are combined. The next feature is the joint use of the same scales for different ribs, which speeds up and simplifies training. For clarity, the convolutional layer scheme can be represented in the form of a filter sliding along the selected axes (there can be 1,2,3 or more). One filter passes through all possible points and forms the next level of neurons by scalar multiplication of the filter values by the values of the object. The number of parameters is equal to the product of the number of filters and the filter size. During the current experiments, two-dimensional convolutional layers were used, in which one axis corresponded to the measurement time and the second to the position of the sensor. Thus, a two-dimensional map of signals is created.

As the advantages of this type of layers it should be noted a small number of parameters that are limited by the size of the filter, the sharing of weights. These factors accelerate the training of the neural network and find specific features of objects along the indicated axes. It is these features that make convolutional neural networks very popular for solving pattern recognition tasks (2 axes - width and height of the image), text analysis (1 axis - letter / word position in the text) or audio (one axis - time, or two axes - time and sound frequency)

Of the disadvantages - the use of this type of layer is limited to certain types of tasks.

c) MaxPooling layers

MaxPooling (also called AveragePooling) layers are also typical for convolutional neural networks. These layers are nonparametric, and their work is to select the maximum (or average) value inside the given window and transfer this value to the neuron of the next layer (**Fig. 5**). Pooling layers allow us to reduce the number of neurons in the next word 4 (9, 16, ...) times, thereby reducing the number of

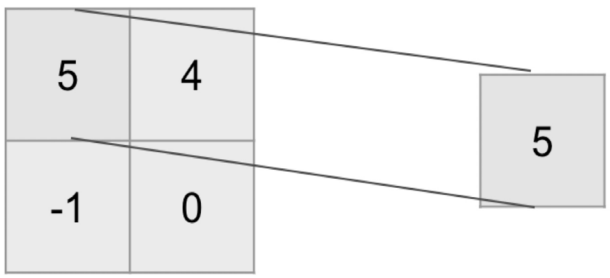


Fig. 5. *MaxPooling layer.*

weights on the next layers. In practice, these types of layers are almost universally used in the training of convolutional neural networks to improve their convergence.

In neural networks, output signals from neurons pass through a non-linear activation function. Examples include hyperbolic tangent, sigmoid, ReLU (Restricted Linear Unit, function of the form $y = \max(0, x)$). Without activation functions, a neural network (or a subset of its layers) would turn into a simple linear function, so the presence of nonlinear activations is an essential component of a neural network. The choice of activation function is left to the discretion of the researcher. In this work, we used the ReLU activation function, the graph of which is shown in **Fig. 6**.

The general view of the neural network used for the above experiments can be described as follows:

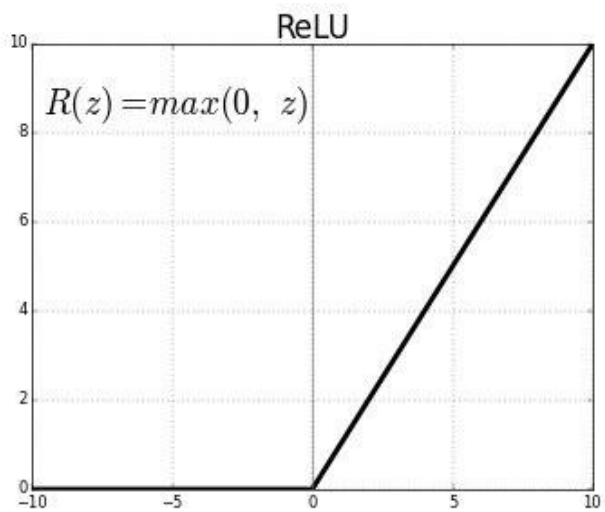


Fig. 6. *The graph of activation function ReLU.*

1. Convolutional layer (32 filters (cores), 3×3 core size, ReLU activation function)
2. MaxPooling layer (size 2×2, without activation function)
3. Convolutional layer (32 cores, 3×3 kernel size, ReLU activation function)
4. MaxPooling layer (size 2×2, without activation function)
5. Fully connected layer (64 neurons, ReLU activation function)
6. The output layer (1,2,3 neurons depending on the task)

The output layer for the problems under consideration was different. In the case of the angle of inclination, thickness, density - one neuron, the angle of inclination and thickness - two neurons, all parameters are three neurons. As a loss function (loss function), the value Mean Squared Error (MSE, root mean square error) was used. During the training, the Mean Absolute Error metric (MAE, average modulo error) was also monitored. These metrics can be used both for a one-dimensional quantity and for multidimensional vectors

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|, \quad MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2.$$

Neural network training

The neural network was trained in 32 mini-batches. The Adam optimization algorithm (adaptive moment estimation) was chosen. This optimization algorithm has many advantages (low memory requirements, simple and computationally efficient). The Adam algorithm is different from the classical stochastic gradient descent methods, which maintain a constant training speed during training to update weights. In the Adam algorithm, the training rate is updated taking into account the estimates of the first and second moment. This algorithm is well established for optimization of regression problems by neural networks. In these experiments, the initial convergence rate was set to 0.001.

3. RESULTS

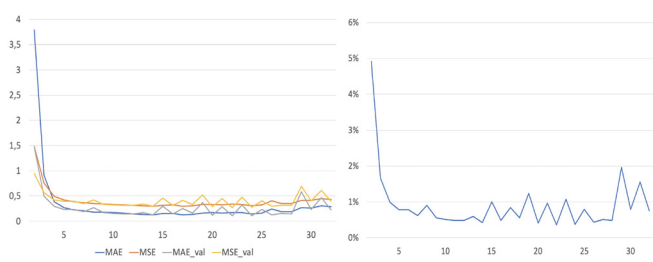
The neural network was trained and validation was carried out on sets of solutions of the direct problem for the tasks with variations: the angle of inclination of the fractures, the height of the fractures, the angle of inclination and height at the same time, the density of the location of the fractures (the distance between them) and all three parameters at the same time.

Angle of inclination

For the problem with a variation in the angle of inclination of fractures in the system in the range from -15 to +15 degrees relative to the vertical (subvertical fractures), training was performed on a set of 4021 solutions of the direct problem and then validation was performed on a set of 1981 control samples. **Fig. 7** shows a graph of the dependence of the average error in recognition on the epoch of training (32 epochs in total). It can be seen that a sufficiently low error is achieved – less than 1%.

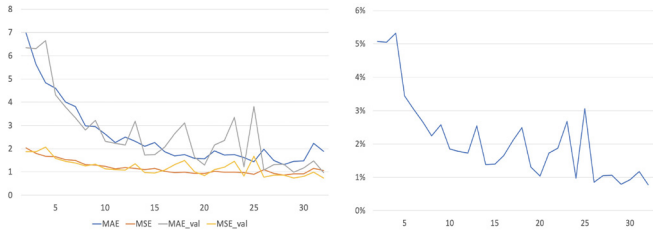
Height of fractures

For a problem with a variation in the height of fractures in a system in the range from 50 to 200 meters, training was performed on a set of 4018 solutions to the direct problem and then validation was performed on a set of 1980 control samples. **Fig. 8** shows a graph of the



a *b*

Fig. 7. *Graphs of the metrics MAE and MSE depending on the number of the training epoch (a) for the training (MAE, MSE) and validation (MAE_val, MSE_val) sets with varying angles. Dependence of the error in determining the angle of inclination of fractures on the number of the epoch of training (b).*



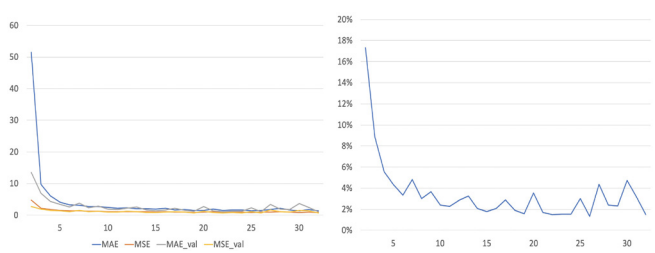
a *b*

Fig. 8. *Graphs of the metrics MAE and MSE depending on the number of the training epoch (a) for the training (MAE, MSE) and validation (MAE_val, MSE_val) sets with varying fracture heights. The dependence of the error in determining the height of fractures on the number of the epoch of training (b).*

dependence of the average error in recognition on the epoch of training (32 epochs in total). It can be seen that a sufficiently low error is achieved - about 1%.

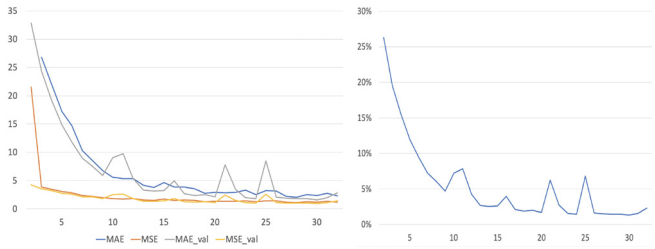
Simultaneous variation in inclination and fracture height

For the problem with simultaneous variation of the angle of inclination of fractures in the system in the range from -15 to +15 degrees relative to the vertical (subvertical fractures) and the height of fractures in the system in the range from 50 to 200 meters, training was conducted on a set of 4020 solutions of the direct problem and then validation was performed on a set of 1981 control samples. The **Fig. 9** shows a graph of the dependence of the average error in recognition on the



a *b*

Fig. 9. *Graphs of the MAE and MSE metrics depending on the number of the training epoch (a) for the training (MAE, MSE) and validation (MAE_val, MSE_val) sets with variation in the angle and fracture height. The dependence of the error in determining the angle of inclination and the height of the fractures on the number of the training epoch (b).*



a *b*

Fig. 10. *Graphs of the MAE and MSE metrics depending on the number of the training epoch (a) for the training (MAE, MSE) and validation (MAE_val, MSE_val) sets with varying fracture densities. Dependence of the error in determining the density of the location of fractures on the number of the training epoch (b).*

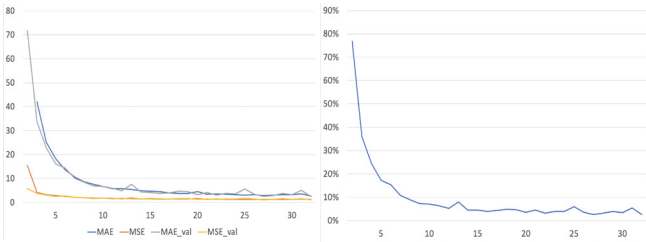
epoch of training (total 32 epochs). It can be seen that a sufficiently low error is achieved - about 3-4%.

Fracture density (the distance between fractures in system)

For a problem with a variation in the density of fractures (distance between fractures) in a system in the range from 50 to 200 meters, training was conducted on a set of 3350 solutions of the direct problem and then validation was performed on a set of 1650 control samples. **Fig. 10** shows a graph of the dependence of the average error in recognition on the epoch of training (a total of 32 epochs). It can be seen that a sufficiently low error is achieved - about 2%.

Simultaneous variation of all three parameters

For a problem with simultaneous variation of the angle of inclination of fractures in the system in the range from -15 to +15 degrees relative to the vertical (subvertical fractures), the height of fractures in the system in the range of 50 to 200 meters and the density of fractures (distance between fractures) in the system in the range of from 50 to 200 meters, training was conducted on a set of 4020 solutions of the direct problem, and then validation was performed on a set of 1981 control samples. **Fig. 11** shows a graph of the



a *b*

Fig. 11. *Graphs of the metrics MAE and MSE depending on the number of the training epoch (a) for the training (MAE, MSE) and validation (MAE_val, MSE_val) sets when all three parameters are varied. The dependence of the error in determining the angle of inclination, height and density of the location of fractures on the number of the epoch of training (b).*

dependence of the average error in recognition on the epoch of training (32 epochs in total). It can be seen that a sufficiently low error is achieved - about 5-7%.

4. CONCLUSION

The research showed good applicability for solving the inverse problems of exploration seismology of macrofractures layers of machine learning techniques, in particular convolutional neural networks, with training sets obtained by solving direct problems by mathematical modeling using the grid-characteristic method. In the most complicated formulation — when varying the three main parameters of a unidirectional fracture system (inclination angle, height and fracture density) — we got a recognition error of only 5-7%, which is a fairly good indicator.

Low error indices make it possible to complicate the task by introducing additional parameters for variation (for example, the spatial position of the system of fractures, as was done for a single fracture in [30]). What will be done in further research.

REFERENCES

1. Sheriff RE, Geldart LP. *Exploration seismology.* Cambridge University Press, 1995, 592 p.
2. Braduchan YuV, Goldberg AV, Gurari FG. *Bazhenovskii gorizont Zapadnoi Sibiri* [Bazhen horizon of West Siberia]. Novosibirsk, Nauka Publ., 1986, 216 p.
3. Dorofeeva EV. *Tektonicheskaya treschinovatost' gornykh porod i usloviya formirovaniya treschinnykh kollektorov nefti i gaza* [Tectonic fracturing of rocks and conditions for the formation of fractured reservoirs of oil and gas]. Moscow, Nedra Publ., 1986, 231 p.
4. Kozlov EA. *Modeli sredy v razvedochnoi seismologii* [Models of medium in exploration seismology]. Tver, GERS Publ., 2006, 480 p.
5. Claerbout JF. Coarse grid calculations of waves in inhomogeneous media with application to delineation of complicated seismic structure. *Geophysics*, 1970, 36(3):407-418.
6. Claerbout JF, Doherty SM. Downward continuation of moveout-corrected seismograms. *Geophysics*, 1972, 37(5):741-768.
7. Etgen J, Gray S, Zhang Y. An overview of depth imaging in exploration geophysics. *Geophysics*, 2009, 74:WCA5-WCA17.
8. Jiao K, Huang W, Vigh D, Kapoor J, Coates R, Starr EW, Cheng X. Elastic migration for improving salt and subsalt imaging and inversion. *SEG Technical Program Expanded Abstracts,* 2012, 1-5.
9. Luo Y, Tromp J, Denel B, Calandra H. 3D coupled acoustic-elastic migration with topography and bathymetry based on spectral-element and adjoint methods. *Geophysics*, 2013, 78(4):S193-S202.
10. Burago NG, Nikitin IS, Yakushev VL. Hybrid Numerical Method with Adaptive Overlapping Meshes for Solving Nonstationary Problems in Continuum Mechanics. *Computational Mathematics and Mathematical Physics*, 2016, 56(6):1065-1074.
11. Nikitin IS, Burago NG, Nikitin AD. Explicit-Implicit schemes for solving the problems of the dynamics of isotropic and anisotropic elastoviscoplastic media. *IOP Conf. Series: Journal of Physics: Conf. Series*, 2019, 1158(3):1-8.
12. Burago NG, Nikitin IS. Algorithms of through calculation for damage processes. *Computer Research and Modeling*, 2018, 10(5):645-666.
13. Fang X, Zheng Y, Fehler MC. Fracture clustering effect on amplitude variation with

offset and azimuth analyses. *Geophysics*, 2017, 82(1):N13-N25.

14. Muratov MV, Petrov IB. Application of fractures mathematical models in exploration seismology problems modeling. *Smart Innovation, Systems and Technologies,* 2019, 120-131.
15. Krizhevsky A, Sustskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012, 1097-1105.
16. Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. *Proc. of the 26th Intern. Conf. on Neural Information Processing Systems* (*NIPS'13*), 2013, 2:2553-2561.
17. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. *Advances in neural information processing systems*, 2014, 2672-2680.
18. Zhang C, Frogner C, Araya-Polo M, Hohl D. Machine-learning Based Automated Fault Detection in Seismic Traces. *EAGE Conference and Exhibition,* 2014.
19. Araya-Polo M, Dahlke T, Frogner C, Zhang C, Poggio T, Hohl D. Automated fault detection without seismic processing. *The leading edge,* 2017, 36(3):194-280.
20. Wu Yu, Lin Yo, Zhou Zh, Delorey A. Seismic-Net: A Deep Densely Connected Neural Network to Detect Seismic Events. *CoRR,* 2018.
21. Menga M, Chua YJ, Woutersonb E, Ong CPK. Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks. *Neurocomputing*, 2017, 257:128-135.
22. Petrov IB, Muratov MV. The application of grid-characteristic method in solution of fractured formations exploration seismology direct problems (review article). *Matem. Mod.*, 2019, 31(4):33-56.
23. Leviant VB, Petrov IB, Muratov MV. Chislennoe modelirovanie volnovikh otklikov ot sistemy (klastera) subvertikalnykh makrotreschin [Numerical simulation of wave responses from subvertical macrofractures system]. *Technologii seismorazvedki*, 2012, 1:5-21 (in Russ.).
24. Novatskii VK. *Teoriya uprugosti* [Elastic theory]. Moscow, Mir Publ., 1975, 872 p.
25. Magomedov KM, Kholodov AS. *Setochno-kharakteristicheskie chislennye metody* [Grid-characteristic numerical methods]. Moscow, Nauka Publ., 1988, 288 p.
26. Petrov IB, Tormasov AG, Kholodov AS. On the use of hybrid grid-characteristic schemes for the numerical solution of three-dimensional problems in the dynamics of a deformable solid. *USSR Computational Mathematics and Mathematical Physics*, 1990, 30(4):191-196.
27. Favorskaya AV, Breus AV, Galitskii BV. Application of the grid-characteristic method to the seismic isolation model. *Smart Innovation, Systems and Technologies*, 2019, 133:167-181.
28. Leviant VB, Petrov IB, Kvasov IE. Chislennoe modelirovanie volnovogo otklika ot subvertikalnykh makrotreschin, veroyatnykh fluidoprovodyaschikh kanalov [Numerical modeling of seismic response from subvertical macrofractures as possible fluid conduids]. *Tekhnologii seismorazvedki*, 2011, 4:41-61 (in Russ.).
29. Leviant VB, Miryakha VA, Muratov MV, Petrov IB. Otsenka vliyaniya na seismicheskii otklik raskrytosti treschiny i doli ploschadi lokalnykh kontaktov k ee poverkhnosti [Seismic responses of vertical fractures depending on their thickness]. *Tekhnologii seismorazvedki,* 2015, 3:16-30 (in Russ.).
30. Muratov MV, Biryukov VA, Petrov IB. Solution of the Fracture Detection Problem by Machine Learning Methods. *Doklady Mathematics*, 2020, 101(2):169-171.

DOI: 10.17725/rensit.2020.12.263

Oculomotor reactions in fixations and saccades with visual perception of information

Rostislav V. Belyaev, Vladimir I. Grachev, Vladimir V. Kolesov
Kotelnikov Institute of Radioengineering and Electronics of RAS, http://www.cplire.ru/
Moscow 125009, Russian Federation
E-mail: belyaev@cplire.ru, grachev@cplire.ru, kvv@cplire.ru

Galina Ya. Menshikova
Lomonosov Moscow State University, http://www.psy.msu.ru/
Moscow 125009, Russian Federation
E-mail: menshikova@psi.msu.ru

Alexander M. Popov, Viktor I. Ryabenkov
MIREA-Russian Technical University, http://www.mirea.ru/
Moscow 119454, Russian Federation
E-mail: popov@mirea.ru, ryabenkov@mirea.ru

Received July 9, 2020, peer reviewed July 13, 2020, accepted July 20, 2020

Abstract. **The work is devoted to the peculiarities of gaze movement in fixations and saccades when reading text and perceiving neutral images. The existence of multiple (almost unidirectional) displacements in fixations was discovered, the probability of which significantly exceeds the similar probability in simulated random fixations. The dependence of the correlation of time spent in fixations on the degree of distortion of texts when reading them was investigated. An interesting aspect of research is the statistics of the distribution of the direction of gaze shifts in fixations and saccades. It is shown that during reading the direction of displacements in saccades is predetermined. In the case of fixations in the distribution, vertical and horizontal directions are distinguished, and this is typical not only when reading a text, but also when perceiving various images, including those rotated at different angles. The individual characteristics of the visual system associated with the non-synchronous movement of the gaze of the left and right eyes in some subjects were found. It was determined the noise error of the eye-tracking recording system when registering the gaze displacements. It is shown that down to gaze displacements of the order of 1 mm, the noises of the registering measurement system prevails, which have a normal distribution along the length, and are uniformly distributed over the angle.**

Keywords: **oculomotor reactions (oculomotorics), technology of eye tracking, fixations, saccades, reading text, multiple shifts**

UDC 004.932.2; 159.931

Acknowledgments: This work was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation.

For citation: Rostislav V. Belyaev, Vladimir I. Grachev, Vladimir V. Kolesov, Galina Ya. Menshikova, Alexander M. Popov, Viktor I. Ryabenkov. Oculomotor reactions in fixations and saccades with visual perception of information. *RENSIT,* 2020, 12(2):263-274. DOI: 10.17725/rensit.2020.12.263.

"Vultus est index animi"
("Eyes are the soul mirror")
Mark Thullius Cicero.

Contents

1. INTRODUCTION

Studies of eye movement in the visual perception of various information are intensively conducted to solve a number of physiological and psychological problems. Each person has a structure of lines, dots and colors in the eye iris combined in unique combinations. Some people may have similar eye color, but the lines and dots on the iris themselves are as unique as fingerprints. It is known that almost everything can be determined at a person's eyes: his mood, character traits, truthfulness of words and many other aspects of his inner world. An analysis of the involuntary eyes reaction to information can tell a lot not only about the individual physiological characteristics of the human visual apparatus, but also about the cognitive and psychological characteristics of the individual, his preferences, positive and negative emotions caused by information, about of hidden and suppressed feelings and emotions.

Recently, another important aspect of research in this area has appeared - the effect of oculomotor reactions on visual acuity [1]. High visual acuity is extremely important in various life situations and for many professional tasks, from confident recognition of objects to driving cars and airplanes. It is well known that the optical and anatomical characteristics of the eye contribute to good vision and spatial resolution, but the effect of reflex eye movements on improving visual acuity has not been studied. Thus, the study of oculomotor activity in the perception of information is quite an urgent task.

2. EQUIPMENT AND RESEARCH METHODS

The eye movement trajectory during visual perception of information consists mainly of local fixations, when the gaze is fixed in the area of individual image elements, and saccades, when the gaze is transferred from one element to another. If the eyes movement in saccades is more less obvious and understandable, then in the fixations the situation is much more complicated, both in terms of the time spent by the gaze in them, and in terms of the gaze trajectory [2]. With visual perception of information of the order of 90% of the time, the gaze is in fixations and, apparently, at this time the formation and cognitive awareness of the visual pattern by the human brain occurs.

The intensity of this activity is far from always the same and strongly depends on the nature of the presented images. For example, at reading of familiar signs (letters, numbers, and other signs) are perceived almost automatically (at a reflex level), it does not require a long examination, and it remains only to understand the text and follow the content of what is read. With a general perception of graphical information (pictures, figures) with neutral content, in the absence of fine details or a masked image in them, the pattern of eye movement is very different from the pattern when reading and consists practically of only saccades.

In this work, eye movement was register using a computer installation with eye tracking technology iView XTM High Speed 1250 IT from the German company SMI GmbH

(resolution < 0.01°, sampling frequency 1250 Hz). In the experiment, the observer's head is located at a certain distance from the monitor screen (80 cm), on which the desired image is presented, and in order to avoid involuntary movements and turns, it is fixed in a special device. The installation works in the "dark pupil" mode. Herewith, the eye is illuminated by a point source of infrared radiation, and the infrared video camera performs high-speed shooting of the eye. In the image, the position of the pupil is determined programmatically (in IR rays it is a dark oval) and its size, as well as the position of the corneal flare, which is a reflection on the cornea of an infrared light source. In the image, the position of the pupil and its size are determined programmatically (in IR rays it is a dark oval), as well as the position of the corneal light patch, which is a reflection on the cornea of an infrared light source. The direction of gaze is calculated based on a vector connecting the positions of the corneal light patch and the center of the pupil. The power level of the IR source is sufficient for experiments, but does not exceed the value dangerous to the eye. The parameters of the infrared eye image after processing by a special computer program are saved in the data file.

Before starting measurements in the calibration process, the deviation of the observer's pupil from its central position is determined on the system of special reference points on the monitor screen. On the developed algorithm, the relationship between the position of the pupil and the position on the monitor screen of the observer's gaze on the image is determined. As a result of using this algorithm, we can track on the monitor screen a gaze at the image, rather than eye movement.

Of the several possible types of eye movement, differing in temporal and spatial characteristics, three basic characteristics of eye movement were registering at the setup when observing the image: saccades, fixations, and the pattern of eye movements formed by the successive selections from fixations and saccades. The trajectories of displaying the movement of the gaze were constructed by connecting the points with time-successive coordinates, determined by the sampling frequency of the digital measuring system.

For the study of involuntary eye movements that are not related to cognitive processes in the brain, two types of gaze displacements are of interest: microsaccades (fixations) - short displacements with a sharp change in direction associated with the process of accommodation of the visual apparatus, and relatively longer saccades associated with gaze transfer to another place and having an average of approximately one direction [4,5].

A pattern of eye movement image in the fixations area is a multiplicity of closely and randomly located points on which the gaze passes sequentially. Herewith a comparative statistical analysis of the points distribution in the fixation region is of interest. A comparative model for the pattern of eye movement image in the fixations region can be the random distribution with specified parameters (size of the fixation region, dispersion and approximately the same points density over the entire fixation area, average displacement). Thus, if using the random number generator we create an artificial "random" fixation (CF) with the necessary parameters, then its statistical characteristics and some parameters can serve as if standards when comparing with similar characteristics of real fixations (RF). The evident differences in these characteristics, that are detected, are likely to be due to cognitive processes in the brain.

The gaze movement in the fixation area is also characterized by the temporary distribution

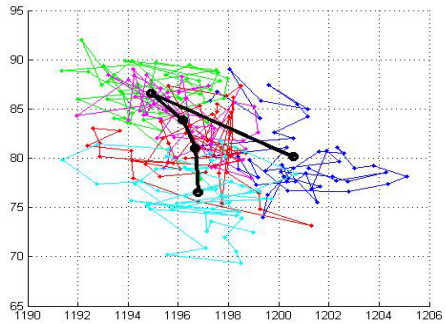


Fig. 1. *Sequential fragments of fixation (shown in color) with a length of 50 points. The centers of gravity of fragments drift along a thick solid line.*

of points in it. If the RF is divided into separate fragments and the position of the center of gravity is determined for each fragment, then connecting them with a solid line, it turns out that the centers of fragments gravity as if drift along the whole fixation, further complicating the creation of the SF model (**Fig. 1**).

From general considerations, the movement of the gaze in fixations should provide the solution of two problems: retention of the image of some moment considered image element in the fovea region (the central area of the retina is the region of the highest density of cones) and ensuring the absence of the effect of receptors saturation [6].

The experiment involved several testees aged 20 to 65 years. Let us conditionally denote them by the letters E, O, V, M, Z, P4, P5 and P7. The first five were asked to view, for no specific task, for 10 seconds, three neutral images **Fig. 1***a,b,c* (we will conditionally denote them by F, T, and W), which at each subsequent observation turned 45° relative to the previous position, moreover, in testees M and Z the

a *b* *c*

Fig. 1*a,b,c. Images of various structures presented in the experiment: a - "fractal" (F); b -"tree" (T) and c - "wave" (W).*

coordinates of both the left and right gaze were registering, but with a frequency $f = 500$ Hz.

Testees P4, P5, and P7 were asked to read 11 short texts (3-4 lines), herewith the first version of each text fully corresponded of orthography, while the reading of the next five versions of the same text was deliberately complicated by distortions: punctuation marks and gaps between words were excluded, extra spaces were inserted, the letters in the words changed places and at the same time the words were rearranged or torn.

3. ANALYSIS OF THE GLANCE MOVEMENTS TRAJECTORY IN THE FIELD FIXATIONS AREA

For the CF model, the probability of implementation of several approximately unidirectional consecutive displacements (hereinafter referred to as multiple displacements, moreover by the displacement multiplicity k, we mean the number of consecutive, unidirectional displacements) should decrease sharply with increasing multiplicity k. The experimental data show that the number of such multiple displacements in real fixations (RF) always exceeds the analogous number for CF at the identical statistics (full number of displacements in fixation). The procedure for detecting multiple displacements in the RF is as follows: when the RF length is equal to N, the sums of the lengths of k successive displacements were calculated (obviously such sums were accumulated $(N\text{-}k\text{+}1)$), then from the each thus obtained sum was deducted the length of displacement from the beginning of the first displacements to end of k-th. If the obtained difference turned out to be less than a tenth of the average displacement length in the RF, then it was believed that the data of k displacements have approximately the same

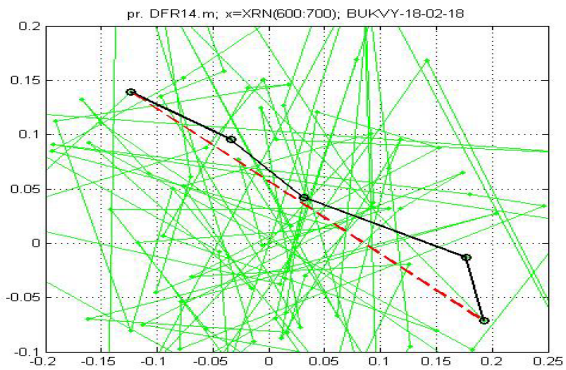


Fig. 2. *A fragment of real tracking, in which 4 consecutive displacements made in approximately the same direction and a dashed line highlighting the movement between the start and end points of the track are highlighted in a bold line.*

direction, otherwise these k displacements were not considered unidirectional.

The described procedure for the multiplicity $k = 4$ is illustrated in **Fig. 2**. In fact, the resulting displacement along k segments of the broken line is compared with the sum of the lengths of the same segments elongated along a straight line (in one direction). Note that one tenth of the average displacement length is of the order of 0.04-0.2 mm and is a rather hard selection criterion.

In total, more than 103 fixations of various lengths were processed. The processing results are presented in graphs (**Fig. 3***a* and **3***b*), showing the ratio of the multiple displacements number to the total displacements number depending on the multiplicity k. In the same axes, a similar

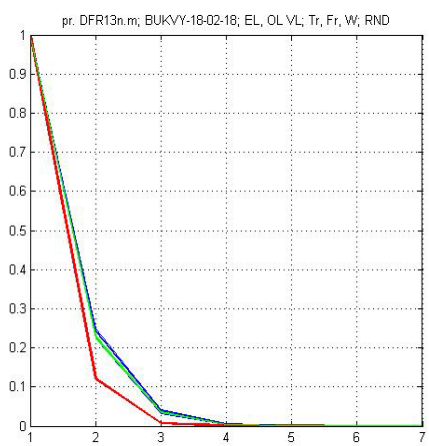


Fig. 3a. *Dependence of the fraction of multiple displacements on the multiplicity k (the red curve for the CF is the rest for the RF).*

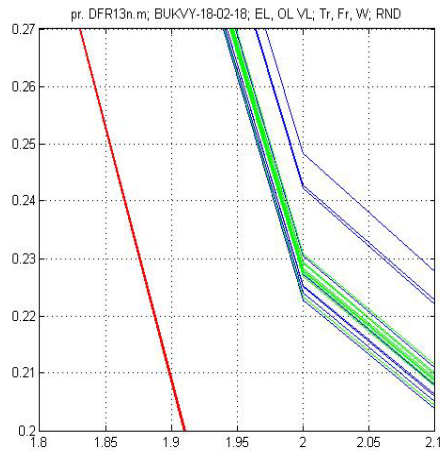


Fig. 3b. *Dependence of the fraction of multiple displacements on the multiplicity k (the red curve for the CF is the rest for the RF).*

relation was constructed for artificial random fixation of the CF, which was generated so that the average displacement length in it was the same as in the RF, and the number of points coincided exactly.

The appearance of the CF and RF graphs is exactly the same, the multiple displacements number in the CF for any multiplicity is always less than in real ones, and for multiplicity $k > 4$ they are completely absent (the fixation contains about 50,000 points), while in real fixations there are displacements, the multiplicity of which reaches 7 even with a shorter duration.

At the same time, a very small scatter in the values (graphs practically merge) of the indicated dependences for real fixations, irrespective of the testee and the object under consideration, attracts attention.

Of particular note is the fact that these dependencies for real fixations have a very small spread in values (the graphs practically merge), regardless of the testee and the object under consideration. Apparently, this indicates that the mechanism (algorithm) of eye movement in fixations is identical in all testees. Given the temporal characteristics of movement, we can assume the presence of a specific video processor (apparently located in

the brain), which should be as close as possible to the muscles performing these movements [6]. Based on the experimental data, a search was carried out for possible preferred directions with multiple displacements in the fixation region. Since we know the starting and ending points of displacements, depending on which one is closer to the center of fixation, one could be judge a certain preferred direction of multiple displacements and draw conclusions about their purpose. However, the analysis of the tracks showed that such a predominant direction does not exist, up to a multiplicity of k = 4, with large multiples a certain difference appears, but it is not provided statistically, the angular distributions of multiple displacements do not reveal any the peculiarities.

4. EYE MOVEMENT ANALYSIS WHEN READING TEXTS

Obviously, when reading texts, the gaze movement is strictly determined. This fact is easily confirmed by numerous gaze tracking, which easily recognizes the movement along the lines, with a certain number of stops in fixations, saccade movement between words and return saccades, returning the gaze to the beginning of the next line [5,6]. Herewith, the gaze spends the same 90% of the time in fixations and, apparently, at this time the reading and preliminary comprehension of what is read takes place. The idea of the experiment described below was to make the look spend more time in fixations, artificially complicating the process of reading and perceiving the text, distorting it in a certain way. How this is done can be illustrated by the example of one of 11 texts. First, we give the source text and 5 distorted versions of it.

1) any point that is estimated to be located at the same distance from the eyes as the fixation point forms two projections of the corresponding points of the retinas,

2) anypointthatisestimatedtobelocatedatthesameistancefromtheeyesasthefixationpointformstwoprojectionsofthecorrespondingpointsoftheretinas,

3) anypointthatisestimatedtobelocatedatthesamedistancefromtheeyesasthefixationpointformstwoprojectionsofthecorrespondingpointsoftheretinas,

4) anyp ointt hati se stimatedt ob el ocateda tt hes amed istancef romt hee yesa st hef ixationp ointf ormst wop rojectionso ft hec orrespondingp oints o ft her etinas,

5) ayn pin ot taht is esemtaidt to eb lotacdet ta hte sema dinsatcne form hte esye sa het faxiniot piotn fomsr wto porjetcoisn fo het cerropsodnign piotsn fo hte rientsa,

6) an y po int th ati ses ti ma te d to bel oca teda tthes am ed ist an cef ro mt hee yesa st he fi xa ti onp oi nt for mst wop ro je cti on sof thec or re spon di ng po in tso fther et in as.

At first glance, the above examples show that the perception of distorted texts is significantly more difficult compared to undistorted ones and, accordingly, the time spent in fixations should increase, because Saccades do not seem to require such an increase in time. Let us denote the ratio of the time spent in fixations to the total time spent reading text through $F_k(i,j)$, where k is the index belonging to a particular testee, i is the text number, j is the variant of its distortion. The calculation results are shown in **Fig. 4***a,b,c*, in the form of corresponding matrices of size (11×6), where each element of the matrix is a ratio of time $F_k(i,j)$, expressed as a percentage.

Unfortunately, to make some unambiguous conclusion from the data in Fig. 4. quite difficult. Only testee F_4 spent the least time reading all eleven versions of the undistorted text. In testees F_5 and F_7, this conclusion was not confirmed.

90.11	90.64	89.81	87.21	92.77	89.61
83.29	89.86	90.24	90.41	89.64	92.38
85.77	91.01	84.70	87.29	87.33	92.86
85.14	92.02	86.74	87.15	90.72	89.22
84.22	93.03	89.60	87.75	90.78	94.35
87.37	90.88	89.79	90.32	90.82	84.28
83.72	89.93	88.33	92.20	88.60	88.53
88.35	93.41	83.54	96.17	91.30	94.13
88.03	87.59	88.95	89.36	86.67	91.99
84.86	92.48	87.19	91.24	90.20	86.97
85.66	89.98	86.13	91.50	90.91	92.55

a) F_4 (*testee no.*4)

98.94	94.54	99.19	100.00	97.48	99.57
99.22	99.40	99.34	92.25	97.31	96.43
98.86	83.48	97.24	92.25	99.28	99.62
98.96	99.50	98.92	95.76	90.69	97.61
99.04	96.86	98.71	94.84	99.18	99.63
86.58	99.39	98.13	98.52	96.66	96.17
98.94	99.44	97.11	97.87	99.28	88.78
99.29	99.47	98.20	98.58	99.07	99.63
98.99	99.26	99.06	99.60	99.30	95.37
94.33	99.34	99.21	97.69	99.43	99.48
99.02	99.38	98.64	98.29	94.39	99.57

b) F_5 (*testee no.* 5)

99.45	99.50	99.38	99.61	99.70	99.62
99.33	99.75	99.53	99.26	99.36	98.39
99.46	99.51	99.61	99.32	99.62	99.52
99.30	99.71	99.53	99.55	99.35	99.73
99.22	99.41	99.35	99.54	99.54	99.74
99.35	99.55	99.08	99.51	99.58	99.51
99.28	99.42	99.35	99.70	99.35	99.52
99.17	99.72	99.59	99.58	99.27	99.65
98.92	99.55	99.24	99.61	99.34	99.60
99.15	99.52	99.03	99.61	99.36	99.68
99.38	99.67	99.36	99.76	99.44	99.65

c) F_7 (*testee no.* 7)

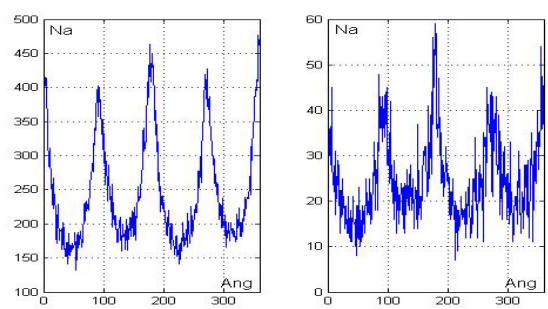
Fig. 4. *Tables of the ratio of the time F_k (i, j) spent in fixations to the total time spent reading the text.*

5. ANGULAR DISTRIBUTIONS OF DISPLACEMENTS IN FIXATIONS AND SAKKADES

Along with the study of approximately unidirectional, sequential displacements of the gaze movement in the fixations and saccades, it is also important to study the angles at which individual displacements are made for times t = 0.8 ms or t = 2.0 ms, depending on the frequency, with which the gaze position was recorded. For each displacement, the angle at which it occurred was calculated, and the angle was counted counterclockwise from the horizontal axis. The angles were calculated separately for displacements belonging to fixations, saccades, or the whole tracking, herewith the belonging of each point to a fixation or saccade was determined in a standard way programmatically and displayed as the corresponding attribute in the data file. Ultimately, angular displacement distributions were constructed in polar or Cartesian coordinates.

Typical distributions in Cartesian coordinates are shown in **Fig. 5**. It should be noted that the size of the fixation region is usually small, its area is about 1–100 mm^2, and, as a rule, the entire fixation is completely projected onto the retina fovea region. In this case, the fixation area is the area of a rectangle located in the plane of the monitor and having an angular size of the order of one degree and the lengths of the sides of which are determined by the extreme coordinates of the fixation points along the corresponding axes.

A certain spread of points belonging to the fixation can determined by the noises of the recording system. But if the noises of the system played an overwhelming role, then the displacements angular distribution in polar coordinates would be an almost strict circle, which is quite easily verified using a random number generator in the corresponding CF model. However, all angular distributions for displacements in real fixations and



fixations *saccades*

Fig. 5. *Integral (total) angular distributions for fixations and saccades (testee E, 8 image positions T). About 200 fixations and saccades were studied.*

saccades show obvious deviations from the circumference, and this statement is true for fixations and saccades of any duration. The characteristic signs of asymmetry are manifested in distributions for virtually every fixation. To a lesser extent, the above statement refers to saccades because of significantly less statistics.

Similar distributions take place for all testees; asymmetry, expressed to a greater or lesser extent is clearly present in all angular distributions.

The perception of rotated images observed by subjects M and Z was also studied, herewith the coordinates of the gaze for the left and right eyes were recorded simultaneously. Corresponding angular distributions are presented in **Fig. 6,** from which it is obvious that the distributions of testee Z are almost the same, while for M they are slightly different, and the difference is manifested for any orientation of the image, which is apparently due to some individual characteristics of the visual apparatus.

It is noteworthy that the asymmetry of the angular distribution of the direction of displacements when observing neutral images for saccades is almost the same as for fixations. A completely different situation is observed when the eye movement is predetermined to be rigidly determined (for example, reading). In **Fig. 7** shows the angular distribution of displacements in all saccades that arise when the

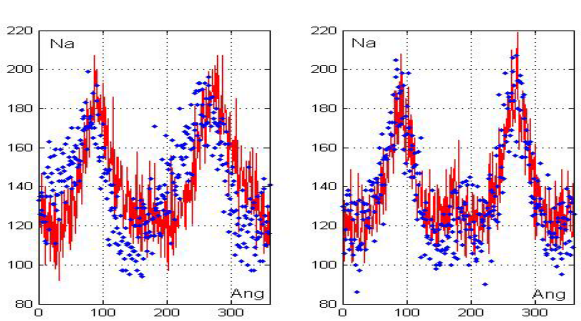


Fig. 6. *Angular distributions for fixations (testee M - on the left, Z - on the right).*

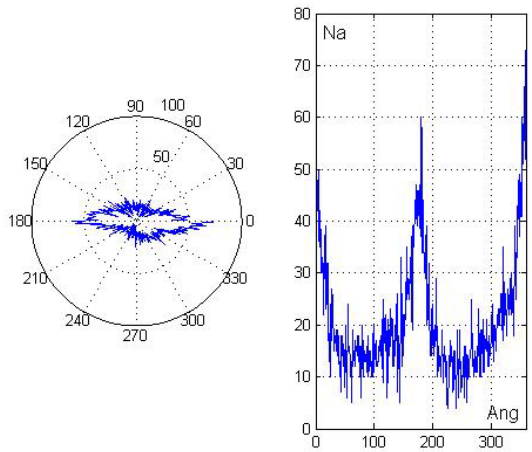


Fig. 7. *Angular distributions for all saccades that arise when the testee P reads six variants of the second text (7598 points in total).*

testee P reads six variants of the second text. The asymmetry clearly changed in a completely predictable way, the angular distribution in the fixations practically remained the same.

We also studied the change in the direction of gaze movement during the implementation of two successive displacements. Obviously, this change is determined by the difference in the angles of directions of the two indicated displacement. f the angles for the displacements vary from 1° to 360°, then the difference of these angles will take values from -359° to + 359°.

For comparison, we consider a model of artificial random fixation (CF), for which the angular distribution of the gaze displacements direction will be uniform, and in polar coordinates it will be a circumference. For such a model fixation, the rotation angles of the displacements are determined as the difference when subtracting the previous angle from the subsequent angle. The angular distribution of such rotation angles (or deviation) is shown in **Fig. 8**. The same distribution is given there for real fixations (RF) of testee E, taking into account all image orientations T (AgW1elT). The number of points in the CF and the RF is taken so that the statistics are the same (88,615 points).

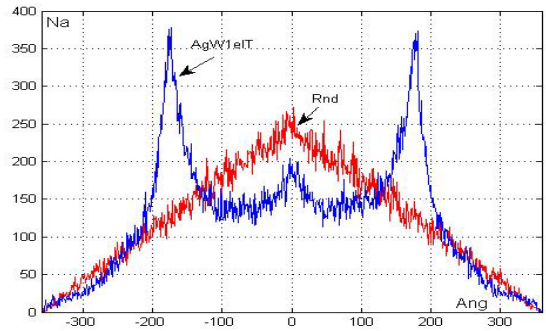


Fig. 8. *Angular distributions for the angles of deviation of the "ideal" fixation (Rnd), for all fixations of the test E taking into account all the orientations of the image T (AgW1elT).*

The fundamental difference between one distribution and another one is obvious.

Moreover, the distribution for "perfect" fixation (CF) remains unchanged, if by the angle of rotation we mean the angle between any pair of displacements. For real fixations (RF), this statement is false, which is shown in **Fig. 9**, which shows the distributions for the rotation angles of displacements whose ordinal numbers differ by 1 (AgW1elT) and 2 (AgW2elT). With increasing difference in ordinal numbers, the distribution strives for Rnd in Fig. 8 for CF. Apparently, such a difference in the distribution of real fixations from "purely random" ones indicates a certain correlation of the nearest two or four gaze displacements, which is completely absent in the random process.

The angular distribution for the displacements in the fixations, presented

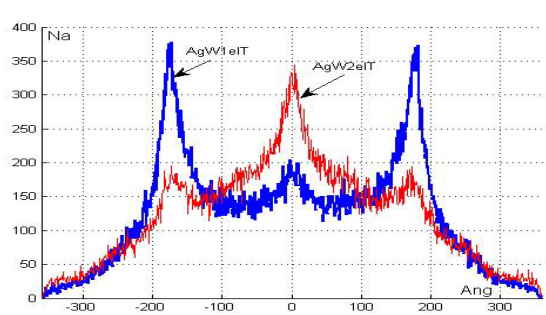


Fig. 9. *Distributions for rotation angles of displacements whose ordinal numbers differ by 1 (AgW1elT) and 2 (AgW2elT).*

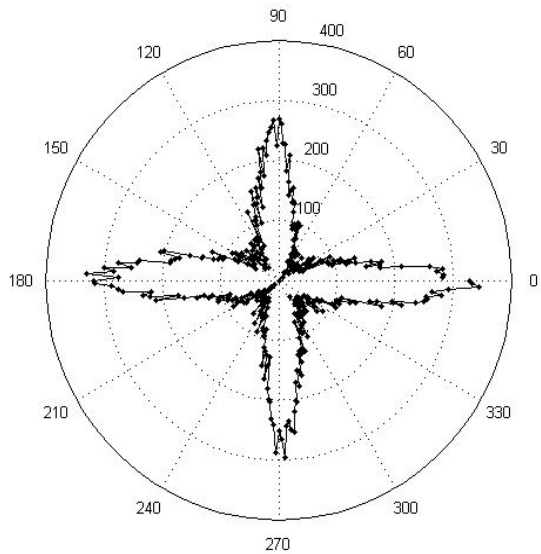


Fig. 10. *The corrected angular distribution of AgFET-132.*

in Fig. 5 includes purely noise offsets, the distribution of which is an almost strict circumference. If we take the radius of such a circumference equal to the minimum value in the full distribution (in the indicated case, this value is 132) and exclude it from the full distribution, then we can obtain the angular distribution of real gaze displacements, which is presented in **Fig. 10**.

The distribution clearly shows the predominance of vertical and horizontal directions in the angular distributions of displacements in fixations and saccades. This is possibly due to the natural coordinate system in the perception of visual information associated with the horizon line and vertical direction (gravity direction).

6. ANALYSIS INFLUENCE OF NOISES OF RECORDING EQUIPMENT

The distribution (Fig. 10) clearly shows the predominance of vertical and horizontal directions in the angular distributions of displacements in fixations and saccades, which is apparently associated with the system of oculomotor muscles. The oculomotor muscles help to carry out the coordinated

movement of the eyeballs, and also, in parallel, provide a high-quality perception of visual information. The motor function of the eye is provided by six muscles - four of them are straight, and two are oblique. Their names are associated with the peculiarities of the course in the eye cavity, where they are located, as well as with the place of attachment to the eyeball wall. Thanks to these muscles, the eyes can perform numerous movements, both unidirectional and multidirectional. Unidirectional turns are up, down, left and others, and multidirectional - bringing the gaze to one point.

When measuring the trajectory of eye movement, there are always noises that are random in magnitude and direction and are superimposed on real gaze displacements. For the registering equipment used, the resolution is 0.01°, and the working accuracy is 0.25 ° - 0.5 ° [10], which in the image plane on the monitor is 0.14, 3.5 and 7.0 mm, respectively (distance to the image is equal 800 mm).

To study the angular distribution of gaze displacements in fixations and saccades, a landscape image with a pronounced horizon line and with a set of vertical elements was used. To reduce the influence of the dominant directions (vertical and horizontal), the image was rotated 315 degrees (**Fig. 11**).

Fig. 11. *Test images to study the angular distribution of gaze displacements in fixations and saccades.*

The division of the trajectory of eye movement into fixed states and saccades was carried out according to known methods [4].

To study statistical characteristics, it is necessary to first perform the following operation: find the coordinates of the "center of gravity" (mean values) for each sampling (or simply the center of the fixation) and subtract the corresponding mean values from each of its coordinates. Having performed such a transfer operation for each fixation and having built a cumulative tracking system, one can get a complete picture of the gaze movement in all fixations when viewing this image. By performing this transfer operation for each fixation and building an aggregate tracking system, one can get a complete picture of the gaze movement in all fixations when viewing that image. Having processed the total set of trackings, we get statistical characteristics, the distribution of points in fixations, for example, along the X and Y axes, corresponding to the variance of distributions, etc. The advantage of this approach is that it is possible to compare the fixations obtained when viewing different patterns by the same testee, and by combining them, one can easily collect the necessary statistics.

Fig. 12*a* shows an aggregate of the set of points for all fixations in one image, plotted relative to a single center of gravity. The trackings registered by the measuring system in fixed states (fixations), corresponding to the gaze displacements in these areas, at least visually look very similar to a random process. In order to compare the results, Fig. 12*b* shows samplings for a true normal process, generated by a random number generator with dispersions in X and Y the same as for the experimentally recorded process.

When observing the trackings of eye movement built for a aggregate of fixations, it is seen that the gaze moves within a limited

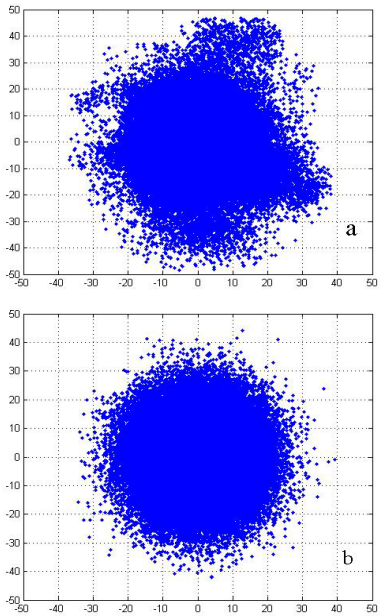


Fig. 12. *a - the aggregate of the set of points in all image fixations, obtained in the experiment, b - modeling of the fixation by random number generator samples.*

area in a chaotic manner (Fig. 12a). It can be seen from the figure that individual points fall out of the specified area, but, due to the large statistics, this practically does not affect the final result.

A typical example of such an angular distribution for the image in Fig. 11 is shown in **Fig. 13**. It can be seen that the angular distribution demonstrates clear asymmetry for all displacements of Ns (green line), as well as for smaller (red line) and larger (black line) values.

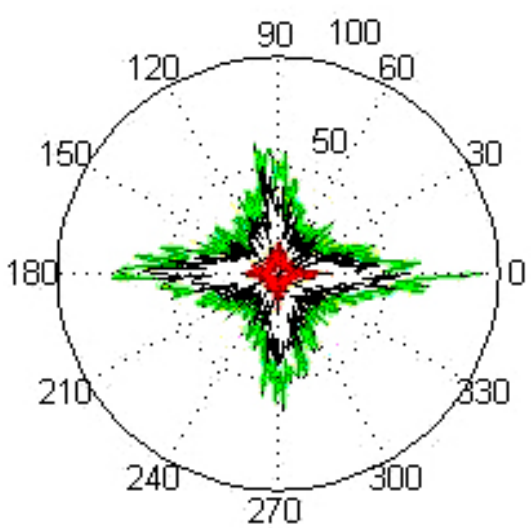


Fig. 13. *Angular distribution of displacements in all fixations for the image in Fig. 11.*

By analyzing the angular distribution for different values of gaze displacements, it is possible to determine the threshold displacement value when the distribution becomes asymmetry. It should be noted that in this case it is necessary yet to exclude "transitional" displacements from fixations to saccades and vice versa, although they are unlikely to have any effect on the summar angular distributions, due to their small number and the absence of any selected direction in the fixation. The resulting distributions are shown in **Fig. 14** at different values of the gaze displacement (LV). The distribution statistics include 267,290 displacements.

Fig. 14 clearly demonstrates that with an increase in the cutoff level (displacements whose length is less than the LV threshold are taken into account), the asymmetry in the angular distribution increases, more and more the horizontal and vertical directions of the gaze diaplacement begin to prevail, and ultimately the angular distribution tends to the shape of a cross (Fig. 13).

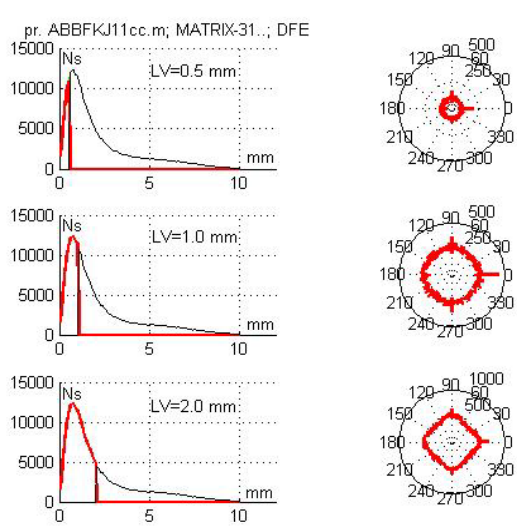


Fig. 14. *Distributions of displacements along the length (left), threshold (0.5, 1.0 and 2.0 mm, vertical line on the graphs on the left), less which the displacements are selected and angular distributions are plotted for them (right).*

Thus, up to gaze displacements LV = 1 mm, we are dealing with noises of the registering measurement system, the maximum of which is approximately 0.8 mm, which have a normal distribution along the length, and are uniformly distributed over the angle, which and observed in Fig. 14. With the growth of displacements, noises in angular statistics plays an ever smaller role, and the role of real of gaze displacements is increasing, as indicated by the angular distribution by changing its shape, deviating more and more from the correct circumference.

7. CONCLUSION

The paper explores some features of the gaze movement in fixations and saccades during visual perception of images and reading texts. It is shown that in the fixations region there are so-called multiple displacements, the frequency of which exceed two or more times than that for model artificial random fixations.

It has been shown that the left and right eye tracking may differ. This is observed both on the trackings itself and on the angular distributions of displacements.

It is shown that the angular distributions of gaze displacements are not uniformly. Vertical and horizontal directions evidently stand out on they. This is characteristic not only for strictly deterministic processes (reading), but remains true and when considering neutral images, including those rotated at different angles.

It was determined the noise error of the eye-tracking recording system when registering the gaze displacements. It is shown that down to gaze displacements of the order of 1 mm, the noises of the registering measurement system prevails, which have a normal distribution along the length, and are uniformly distributed over the angle.

REFERENCES

1. Janis Intoy, Michele Rucci, Finely tuned eye movements enhance visual acuity. *Nature Communications,* 2020, 11:795, doi: 10.1038/s41467-020-14616-2, www.nature.com/naturecommunications.
2. Menshikova GYa, Belyaev RV, Kolesov VV, Ryabenkov VI. The evaluation of individual differences using fractal analysis of scanpaths. *Abstract Book of 18-th Europien Conference on Eye Movements (ECEM*), 16-21 Aug. 2015, Vienna, Austria; http://www.jemr.org/online/8/4/1.
3. Guestrin ED, Eizenman M., General Theory of Remote Gase Estimation Using the Pupil Center and Corneal Reflections. *IEEE Transitions on biomedical engineering*, 2006, 53(6):1124-33.
4. Salvucci D, Goldberg J. Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Eye Tracking Research and Applications Symposium*, New York, ACM Press, 2000.
5. Rostislav V. Belyaev, Vladimir V. Kolesov, Galina Ya. Men'shikova, Alexander M. Popov, Viktor I. Ryabenkov. Study of the special features of the perception of videoinformation by the fractal analysis of the traectory of the eyes motion. *Radioelectronics. Nanosystems. Information Technologies* (*RENSIT*), 2011, 3(1):56-68.
6. Kirpichnikov AP. Eye video processor: Micromovements as a factor of accommodation and video processing. *12th International Conference "Digital Signal Processing and Its Application, DSPA-2010, Image Processing and Transmission"*, Moscow-2010, pp.170-172.
7. Rostislav V. Belyaev, Vladimir V. Kolesov, Galina Ya. Menshikova, Alexander M. Popov, Viktor I. Ryabenkov. Dynamics of reverse saccades of reading texts and its connection with the peculiarities of eye movement. *RENSIT*, 2018,10(1):117-127, DOI: 10.17725/rensit.2018.10.117.
8. Rostislav V. Belyaev, Vladimir V. Kolesov, Galina Ya. Men'shikova, Alexander M. Popov, Viktor I. Ryabenkov. Quantitative criterion of individual differences of the eyes movement trajectories. *RENSIT,* 2015, 7(1):25-33.
9. Tom Foulsham, Alan Kingstone. Asymmetries in the direction of saccades during perception of scenes and fractals: Effects of image type and image features. *Vision Research*, 2010, 50:779-795.
10. Barabanschikov VA, Zhegallo AV. Eye-Tracking Methods: Theory and Practice. *Psychological Science and Education,* 2010, 2(5). (psyedu.ru, ISSN: 2587-6139).

DOI: 10.17725/rensit.2020.12.275

Synchronization Systems Modeling for IEEE 802.11ah Receiver in MATLAB

Fedor B. Serkin, Anton Yu. Dubrovko
Moscow Aviation Institute, https://mai.ru/
Moscow 125993, Russian Federation
E-mail: serkinfb@list.ru, dubrovkoay@mai.ru
Received December 6, 2019, after finalization on July 01, 2020, adopted on July 06, 2020

Abstract. **The problem of applicability of signal processing algorithms that was used in previous Wi-Fi standards to IEEE 802.11ah is discussed. The goal of the work was to study performance of the synchronization algorithms on a low signal to noise ratios. A computer simulation was carried out, that models packet reception in a channel with white noise. Failure probabilities of timing and frequency synchronization systems was measured, as well as bit-error rate. In a model for frequency and timing systems in coarse estimation was used autocorrelation method, in fine estimation – cross-correlation method; for equalizer least-square method was applied; for tracking system two method were explored: classical pilot based and time-frequency decision feedback loop (TF-DFL). As a result, it was confirmed, that TF-DFL method show better results, than classical pilot based one even for traveling pilots. Moreover, in order to approach the theoretical dependence of bit error rate for 10-th modulation coding scheme (MCS10), it is necessary to improve reviewed frequency synchronization and fine timing systems, as well as performance of equalization and tracking methods.**

Keywords: **IEEE 802.11ah, Wi-Fi HaLow, PHY, MCS10, Low SNR, Synchronization, Failure Probability, RFO Tracking, CFO Tracking, Bit Error Rate**

PACS: 84.40.Ua

For citation: Fedor B. Serkin, Anton Yu. Dubrovko. Synchronization Systems Modeling for IEEE 802.11ah Receiver in MATLAB. *RENSIT,* 2020, 12(2):275-286. DOI: 10.17725/rensit.2020.12.275.

Contents

1. INTRODUCTION

The IEEE 802.11ah standard was developed by Institute of Electrical and Electronics Engineers to support various Internet of Things (IoT) and Machine-to-Machine (M2M) applications. Its main features are low energy consumption, number of stations (STA) connected to a single access point (AP) is up to 8192 and radio connection distance up to 1 km [1]. This connection distance is possible because of sub-1GHz band operation and usage of Modulation Coding Scheme 10 (MCS10). This scheme arouse interest since it allows synchronization systems operation

with negative values of signal-to-noise ratio (SNR).

For now, there is one publication about synchronization systems for IEEE 802.11ah receiver [2]. In this paper algorithms of coarse and fine synchronization as well as residual frequency and phase offset (RFO) compensation are outlined. There is also improved channel estimation method is presented with SIG field usage. Preamble synchronization algorithms which were used in the paper is similar to algorithms for previous Wi-Fi standards (a/g/n/ac) [3], after all, ah is very close to them. At the same time, we did not reproduce productivity of RFO compensation system for MCS0, and MCS10 there was not examined at all. That is why systems from articles [4, 5] were chosen.

In the first case [4] was proposed blind estimation method for schemes with modulations from QPSK and higher. The method uses time-frequency decision feedback loop (TF-DFL). It gives quite good performance in SNR range from 4 dB and higher since relation of bit error rate (BER) vs SNR loses around 0.15 dB to theoretical limit. In the same article authors show relation of BER for the most common method which requires usage of pilot sequences. In this case lose to theory is about 1 dB, but in the same time this approach gives better performance with SNR bellow 4 dB. A description is given in the paper [5].

In our work synchronization algorithms performance for IEEE 802.11ah receiver with additive white gaussian noise (AWGN) channel with low values of SNR is considered that were used in the previous Wi-Fi standards.

2. MATERIALS AND METHODS

2.1. Standard

There are many various applications provided by IEEE 802.11ah [6] from smart grids, security systems and targeted advertising to surveillance systems, increasing connection distance to existing hotspots and outdoor Wi-Fi to offload traffic in cell networks. Some applications require coverage of large territories by means of single AP.

Typical IEEE 802.11ah network architecture is shown at **Fig. 1** [1]. This architecture is a centralized network that contains root AP, Relays and STAs. The Relay consist of Relay STA, relay function and Relay AP. At Fig. 1 Relay STAs of Relay 1 and 2 are connected to root AP. Relay STA of Relay 3 is connected to Relay AP of Relay 1. STA 1, which is not AP, is connected to Relay AP of Relay 1. Similarly, STAs 2 and 3 are connected to Relay AP of Relay 3, as well as STAs 4 and 5 are connected to Relay AP of Relay 2. To transmit a frame, for example, from STA 1 to Root AP, it should use relay function from Relay AP to Relay STA of Relay 1. The same but inverse way is passed by frames from Root AP to STA 1.

To reduce latencies in this architecture Transmission Opportunity (TXOP) mechanism is provided. Meanwhile, mandatory access method is Enhanced Distribution Channel Access (EDCA). Besides, there are two optional access

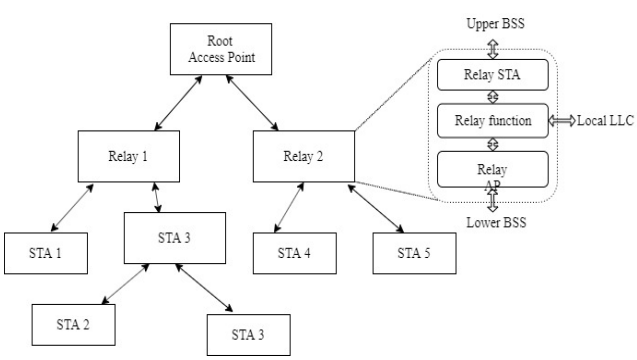


Fig. 1. *Relay network architecture in IEEE 802.11ah.*

methods: Restricted Access Window (RAW), Target Wake Time (TWT). The first one allows to reduce channel load by dividing devices into groups. Within this group there is distribution limited time windows for packet transmission. The second method allows to control packet transmission time from each STA. More information about these methods can be found in standard [1] or in papers [7,8].

On PHY standard provides data transmission with rate from 150 kbit/s over a distance of up to 1 km. This result is achieved by usage of unlicensed sub-GHz band and MCS10. The key parameters of MCS10 and the closest to it schemes that were used in our work are shown in **Table 1.** Papers [8, 9] and document [11] provide a detailed calculation of the radio link budget for our case. We are also interested in what kind of losses on implementation and fading are allowed. According to [12], in previous Wi-Fi standards, these losses accounted for about 12 dB. These losses are calculated using the following formula:

$M[\text{dB}] = RecSens - 30 - P_{MIN}$,

where $RecSens$ [dBm] – minimum receive

Table 1

IEEE 802.11ah parameters of MCS10/0/1

Parameter	Designation	MCS10	MCS0	MCS1
Receiver sensitivity	RecSens [dBm]	-98	-95	-92
Sampling frequency	$1/T_s$ [MHz]	1		
Sampling period	T_s [μs]	1		
FFT length	N_{fft}	32		
FFT period	T_{fft} [μs]	32		
Distance between subcarriers	$\Delta f = \frac{1}{N_{fft}T_s}$ [kHz]	31.25		
Number of subcarriers modulated	N_{dsc}	24		
Number of pilot subcarriers	N_p	2		
Cyclic prefix duration	T_{gi} [μs]	8		
OFDM symbol duration	T_F [μs]	40		
Relative code rate	R	1/2		
Code constraint length	k	7		
Modulation	-	BPSK		QPSK
Generating polynomial	-	[133 171]		

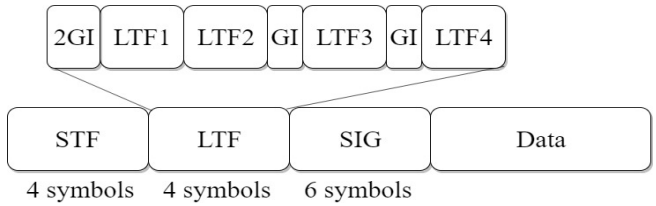


Fig. 2. *S1G_1M packet format.*

sensitivity at which packet error rate (PER) is equal to 10%, P_{MIN}[дБ] = $SNR + 10\lg kT_0W + NF$ receiver sensitivity, SNR – minimum theoretical signal-to-noise ratio at which PER = 10%, for example, for MCS10 it is closely equal to –3.23 dB, [J/K] – Boltzmann constant, T_0 = 293 [K] – the standard noise temperature, W [Hz] – bandwidth, a NF [дБ] – receiver noise figure.

When calculating the radio link budget, the ah working group set the noise coefficient value NF = 7 dB [11]. Fading and implementation loses for MCS10 are equal to the same 12 dB with the corresponding receiver sensitivity.

The main features of MCS10 are: x2 data repetition on half of the subcarriers, short training field (STF) 3 dB gain (**Fig. 2**). Due to repetition relation of BER vs SNR for MCS10 shifts to the negative side by 3 dB relative to MCS0, since the energy of the transmitted symbols is doubled. Thus, an operation area of MCS10 lies in the range of negative SNR. **Fig. 3** shows the shift of the

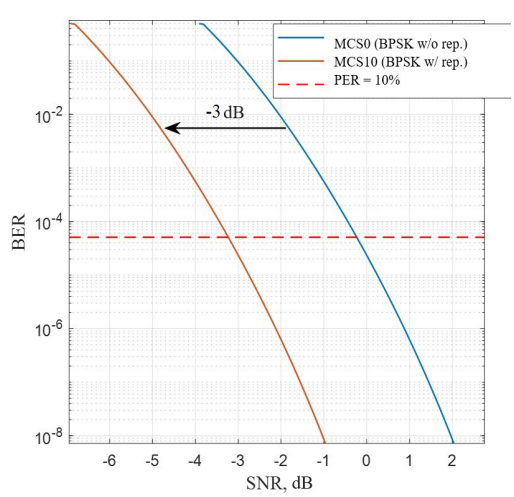


Fig. 3. *Repetition result in MCS10.*

BER curve in this case. Packets transmitted in MCS10 mode have S1G_1M format. Packet structure in this format is shown in Fig. 2.

STF is the short training field, which is used by an automatic gain control system and a coarse synchronization system. LTF is a long training field, which is used by fine synchronization system and equalizer. GI means guard interval. SIG is a signal field, which carries service information about the packet. Data is a data field, which contains a payload.

Data transmission on orthogonal subcarriers **Fig. 4** in the time domain can be described

$$g_i(t) = e^{\frac{j2\pi it}{T_{fft}}},$$

where T_{fft} – FFT duration, i – index of the corresponding subcarrier.

Orthogonality means that the following condition is met

$$\int_0^{T_{fft}} g_i(t) g_l(t) dt = 0, \text{ for } i \neq l.$$

In particular, in standard IEEE 802.11ah to transmit some symbols sequence $c = \{c_1, c_2, ..., c_N\}$ by means of OFDM modulation, it should be split into blocks of M symbols, M = 6 for MCS10.

Next, scrambling is performed. The scrambler structure is shown in **Fig. 5,** under delay blocks their default states are written.

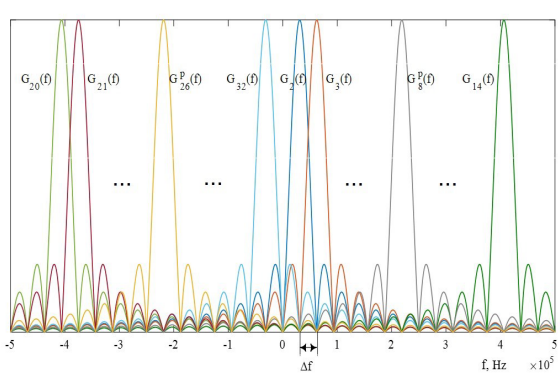


Fig. 4. *Orthogonal subcarriers in MCS10.*

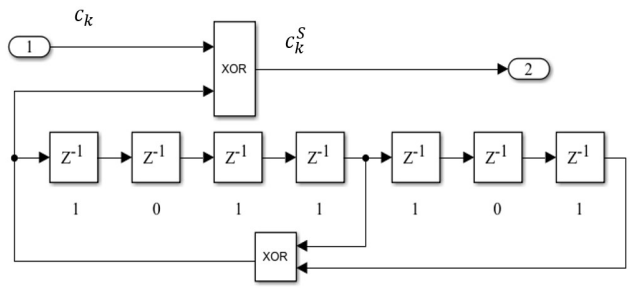


Fig. 5. *Scrambler structure.*

Scrambler's output is c^S.

Now sequence c^S goes to the convolutional encoder, the parameters of which are given in Table 1. The result is a block of 12-bits c'. Then the repetition is performed. The essence of the operation is to copy blocks of 12-bits. The copy is added modulo two with the sequence s = [1 0 0 0 0 1 0 1 0 1 1 1] and then the result is concatenated with block c', thus getting 24-bits block c^{x2}.

Afterward goes interleaving, which allows reducing the probability of correlated multiple errors. An interleaver has a size of 8 rows and 3 columns.

Then the blocks of symbols are converted by BPSK baseband modulator into modulating sequences. To get an OFDM block, pilot sequences and protective zeros are added into modulating sequence. Then IFFT for each block is performed, thus modulating the orthogonal subcarriers. In discrete-time, this looks like

$$s[n] = \sum_{k=0}^{N_{fft}-1} c_k e^{\frac{j2\pi kn}{N_{fft}}}, \ 0 \le n < N_{fft}$$

where N_{fft} is FFT size.

A CP consisting of the last 8 samples of the current OFDM block is added to its beginning, so the OFDM symbol is obtained. After that, windowing is performed that smooth the amplitude of nearby samples at the OFDM symbols borders.

Then digital-to-analog conversion is performed. After interpolation, the signal in

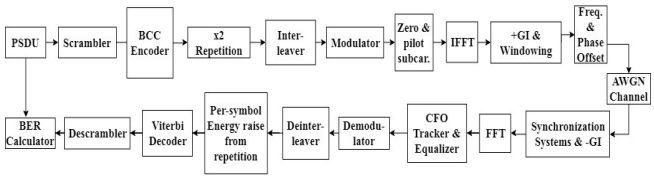


Fig. 6. *Model block diagram.*

continuous time looks like

$$s(t) = \sum_{k=0}^{N_{fft}-1} c_k e^{\frac{j2\pi kt}{T_{fft}}} = \sum_{k=0}^{N_{fft}-1} c_k g_k(t), 0 \le t < T_{fft}.$$

OFDM symbol creation process is shown in **Fig. 6**.

On the receiver side, the signal after being transferred from the carrier, amplified and analog-to-digital converted looks like

$$r[n] = s[n]e^{j2\pi f_{ot} n T_s + \theta_{ot}} + w[n],$$

where $w[n]$ – n-th sample of WGN, f_{ot} and θ_{ot} – some frequency and phase offsets.

2.2. Simulation

A simulation was performed using the synchronization systems described below, as well as equalizer with an estimation of the channel transfer function using the least square (LS) method for the case of AWGN channel with random frequency shifts (Fig. 5). MCS10 packets containing 256 octets of information were transmitted. Packets that did not pass SIG cyclic redundancy check (CRC) were not counted. That is how synchronization failures were filtered out. Frequency offset simulates effect of frequency instability in the transmitter and receiver heterodynes. The standard sets the limit of frequency instability of the transmitter's heterodyne equal to ±20 ppm. Thus, for a carrier frequency equal to 1 GHz, it turns out that the offset after receiver frequency conversion should not exceed ±40 KHz.

To compare our model with other digital data transmission systems the following relation between an energy per bit per noise power spectral density and SNR is fair

$$SNR = EbNo + 10\lg\frac{N_{dsc} + N_p}{N_{fft}} + 10\lg R$$

where $SNR = P_s/N_w$, P_s – the power of a received signal, N_w – the power of WGN, $EbNo$ – the energy per bit per noise power spectral density, N_{dsc} – number of data subcarriers, N_p – number of pilot subcarriers, N_{fft} – FFT size, R – code rate.

The formula for converting BER to the PER has the following form [9]:

$$PER = 1 - (1 - BER)^L,$$

where L – number of transmitted information bits in a packet. This formula is valid only if the errors are independent of each other.

The sample size is calculated based on the following ratio [14]:

$$N = t_\varphi{}^2 \frac{1-p}{\varepsilon_0{}^2 p},$$

where $t_\varphi = 1.96$ is the gaussian distribution quantile for significance level equal to 5%, N – the size of sample bits, ε_0 – related precision of estimation, p – the desired probability.

Incorrect operation criteria of the synchronization system have been adopted: for coarse time synchronization – case when an estimate is out of range of values $t \in [160; 208]$ μs, for fine timing – a difference of estimated value from the true value with some coarse timing estimate, for coarse frequency synchronization – case when an estimate is out of range $\Delta\overline{f}_{coarse} \in [-31.25; 31.25]$ kHz, for coarse and fine frequency synchronization together – case when an estimate is out of range $\Delta\overline{f}_{coarse} + \Delta\overline{f}_{fine} \in [-2; 2]$ kHz.

2.3. Systems

The synchronization system provides OFDM symbols selection at the moment when they begin, as well as performs frequency offset compensation. The offset occurs cause of both transmitter and receiver heterodynes

detuning and the Doppler shift.

The synchronization system divides into time and frequency systems, and each of them also into coarse and fine. Besides, there is a time synchronization up to an integer number of samples and up to fractional one. Fractional time synchronization is not considered in this paper. Now let us look at them separately.

2.3.1. Coarse timing

The coarse frequency synchronization is performed over the STF of the received packet. To do this, the autocorrelation of the signal [3] is calculated using the formula

$$R[n] = \sum_{i=0}^{L-1} r^*[n+i] r[n+i+M],$$

where ( )* - complex conjugate, L – the size of a sliding window, it influences on a value of autocorrelation estimate averaging. In this model L = 80, i.e. a half of STF length. M = 8 is a period of elementary STF sequence. Value M should be multiple of 8, although, with its increase, the delay at the autocorrelator output also increases. By varying the value of M, you can get the peak of autocorrelation at the start of LTF.

Furthermore, a time point is looked for, which corresponds to this maximum, through comparison with some level [3] or by differencing and finding a zero [2]. In this work, the second method was used, since, for the first one, the problem of determining the level arises for small values of SNR. Otherwise, the algorithm's accuracy will be unsatisfactory.

Differentiation is performed using a differentiator, which can be described by equation

$$D[n] = \sum_{i=0}^{N-1} R[n+i] - R[n+i-1],$$

where N – averaging window length, increasing this value leads to the coarse timing estimate with less noise in it, but as a payment, it also leads to some mean bias, i.e. coarse synchronization will be delayed on a corresponding number of samples. In the model N = 32.

The coarse timing is influenced by a moment when the packet was detected since its algorithm depends on the number of STF elementary sequences that are involved in the autocorrelation estimation. As the essence of the algorithm lays a conclusion that the beginning of autocorrelation estimation is located in some interval of values relative to the real beginning, it allows us to achieve a clear peak in LTF beginning moment, which means a more accurate estimate.

2.3.2. Coarse frequency synchronization

The average value of the STF autocorrelation phase was used as an initial or coarse estimate of the frequency offset

$$\Delta\overline{f}_{coarse} = \frac{\overline{\angle R[n]}}{2\pi T_c},$$

where $\angle$ – complex argument, $T_c = TM$ – elementary STF sequence duration, M = 8 – number of samples in elementary STF sequence.

The more elementary sequences are incorporated in calculations, i.e. the faster AGC and detection systems, the more accurate this estimate.

2.3.3. Fine timing

Before demodulation of OFDM symbol data part, it requires to know exact moment of its beginning. That's why after receiving a signal of LTF beginning performs capture of the next 32 samples, then FFT is performed in order to obtain frequency domain signal. In the frequency domain a product of this result $\widehat{LTS}_i$ and reference conjugated LTS. The result is transferred back to the time

domain and complex modulus is taken.

$$xcorr_i = | IFFT(FFT\{\widehat{LTS_i}\}FFT\{LTS\}^*) |,$$

where $i = 1,2,\ldots$ – number of current LTS.

Now maximum sample is in search and its argument is fine timing estimate

$$\hat{t}_{fine_i} = \arg(\max\{xcorr_i\}),$$

where arg( ) – argument of real function.

This value is a number of samples, which should be skipped to start capturing the payload.

2.3.4. Fine frequency synchronization

When OFDM symbol payload is correctly selected from the total sequence, you can refine the coarse frequency estimate. For this mean phase difference between maximum samples of cross-correlation of close LTF symbols is computed, i.e. the phase shift during one OFDM symbol caused by frequency offset.

$$\Delta\hat{f}_{fine_i} = \frac{\angle\max(xcorr_{i-1}) - \angle\max(xcorr_i)}{2\pi T_F},$$

$$\Delta\overline{f}_{fine} = \sum_{i=1}^{3}\Delta\hat{f}_{fine_i}.$$

As a result of refining the variance of aggregated estimate (sum of coarse and fine) significantly reduced (Fig. 11).

2.3.5. Frequency offset tracking

The presence of a frequency offset caused by inaccuracy of estimates of preamble-based synchronization systems described higher due to the presence of AWGN strongly affects the performance of the receiver, so it has to be compensated. For this purpose, tracking algorithms are used. Algorithms from works [4, 5] were studied.

Primarily we consider the most common pilots tracking method [5] without AWGN influence to avoid overloading of mathematics. Pilot subcarriers are averaged and phase is estimated

$$\hat{\varphi}_i = \arctan\left(\frac{\mathrm{Im}\{P_{i,7}\hat{p}_{i,7} + P_{i,-7}\hat{p}_{i,-7}\}}{\mathrm{Re}\{P_{i,7}\hat{p}_{i,7} + P_{i,-7}\hat{p}_{i,-7}\}}\right),$$

where i – index of the current OFDM symbol, P_7 – element of scrambling sequence for 7th subcarrier, $\hat{p}_7$ - received element pilot sequence for 7th subcarrier.

After estimation goes compensation of data vector

$$\hat{D}_i = D_i e^{-j\varphi_i},$$

where $D_i = \{d_{i,1}, d_{i,2}, \ldots d_{i,24}\}$ – data vector, a $d_{i,k} = c_{i,k}e^{j\varphi_i}$ – k-th data subcarrier of i-th OFDM symbol.

Now let us look on the method from [4], so called Time Frequency – Decision Feedback Loop (TF-DFL). In it instead of pilot subcarrier all non-zero subcarriers are used. This method (**Fig. 7**) uses loop that is for a phase estimation is locked in a frequency domain and for a frequency estimation is locked in a time domain.

First of all, for data vector D_i hard decision demodulation is performed. Thess estimates $C_i' = \{c'_{i,1}, c'_{i,2}, \ldots, c'_{i,24}\}$ are used to remove data from data vector, as a result only complex shift is remained for each subcarrier

$$A_i = \{e^{j\psi_{i,1}}, e^{j\psi_{i,2}}, \ldots, e^{j\psi_{i,24}}\}.$$

Then their phase is estimated

$$\varphi_i = \arctan\left(\frac{\mathrm{Im}\{A_i\}}{\mathrm{Re}\{A_i\}}\right).$$

These phases are averaged

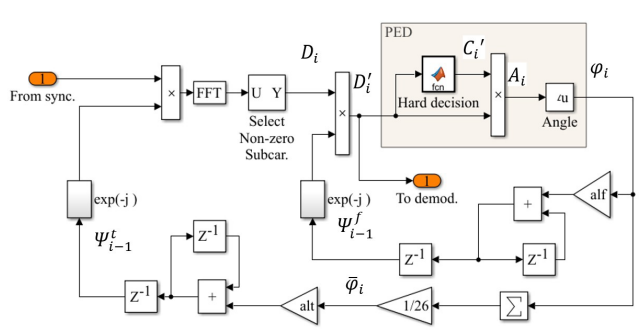


Fig. 7. *TF-DFL block diagram.*

$$\overline{\varphi}_i = \frac{1}{26}\sum_{k=1}^{26}\varphi_{i,k},$$

and the result is multiplied by time loop parameter *alt* = 0.12. Then after integration compensation vector $e^{-j\Psi_{i-1}^{t}}$ is produced. At the end it is multiplied on a current OFDM block in time domain.

To estimate residual phase offset, phases are multiplied by frequency loop parameter *alf* = 0.012. From the result compensation vector $e^{-j\Psi_{i-1}^{f}}$ is produced by integration which corrects current subcarriers in frequency domain

$$D_i' = D_i e^{-j\Psi_{i-1}^{f}}.$$

Reproduced results of TF-DFL tracking from the paper [4] are shown in **Fig. 8**, Also on the same figure BER curve for pilots-based tracking algorithm [5] is presented. This results were obtained with 802.11ah signal in MCS1 mode without coding and with bandwidth of 1 MHz.

2.3.6. Equalizer

In addition to synchronization systems, the equalizer was used in several measurements. It measured channel transfer function with the LS method [15]. This type of equalizer was chosen for its ease in implementation.

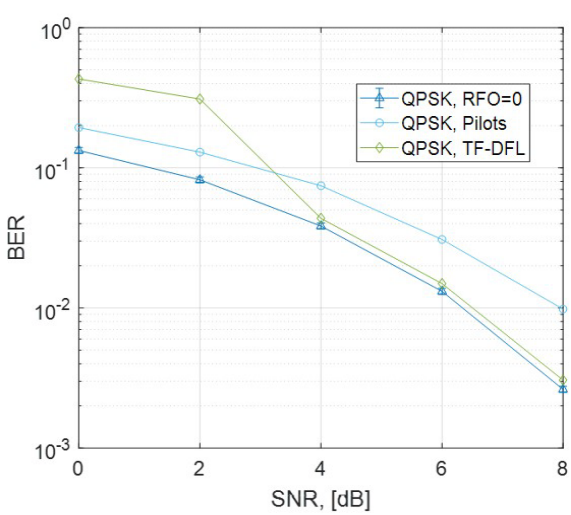


Fig. 8. *Reproduced BER curves from* [4].

Besides, there was no multipath propagation in the model, which means that LS equalizer should not lose in performance against its optimal analog. The estimation was performed using synchronized LTF symbols. The estimates were averaged, based on the assumption that the channel change on the packet duration is insignificant. This process is described in detail below.

Let the spectral density (SD) of the received signal for the *k*-th subcarrier be

$$R_f'[k] = R_f[k]H_f[k] + W_f[k],$$

where R_f[k], W_f[k], H_f[k] – accordingly, the SD of the transmitted signal, the SD of AWGN, the channel transfer function for *k*-th subcarrier.

To estimate transfer function with LS method for *i*-th OFDM symbol on *k*-th subcarrier the following equation is used

$$\widehat{H}_f^{\,i}[k] = \frac{R_f'^{i}[k]}{R_f^{i}[k]}.$$

Then the 4 estimates of the LTF are averaged

$$\overline{H}_f[k] = \frac{1}{4}\sum_{i=1}^{4}\widehat{H}_f^{\,i}[k].$$

Finally, the signal is equalized in the frequency domain

$$R_f''[k] = \frac{R_f'[k]}{\overline{H}_f[k]}.$$

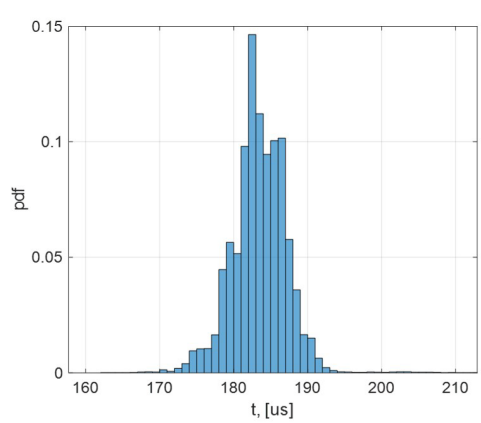


Fig. 9. *PDF of coarse timing.*

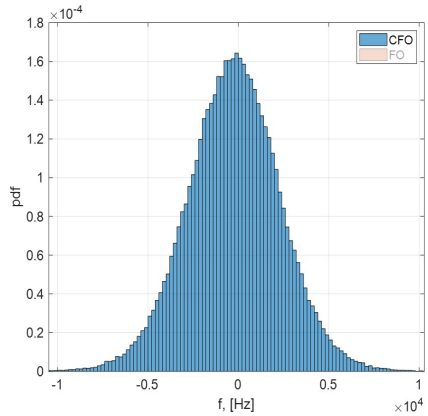


Fig. 10. *PDF of coarse frequency estimates.*

3. RESULTS

This section presents measurements of synchronization system performance for a sample size of 180000 packets with SNR = –2 dB, random phase, and frequency offset (uniform distribution in range ±π and ±40 kHz). Along with BER measurements which are done for a significance level of 5% and relative precision of 10%.

In **Fig. 9** coarse timing histogram is presented.

In **Fig. 10** and **11** coarse and a sum of coarse and fine frequency offset estimates histograms are presented.

In **Table 2** probabilities of preamble-based algorithms failures are presented as well as relative precisions ε0 of these probabilities and confidence intervals for a significance level of 5%. This table shows that the main contribution to the total probability is made by the fine timing and fine frequency synchronization. Based on this fact, the number of packets was limited to the number, which gives close to 10% precision in measuring the probability of failure of these systems.

Table 2

Failure probabilities for timing and frequency synchronization

System	P_{ow}	ε_0	ε_∂
Coarse timing	3.333e-4	0.253	8.433e-5
Fine timing	2.400e-3	0.094	2.260e-4
Coarse frequency sync	1.667e-5	1.131	1.186e-5
Fine frequency sync	2.300e-3	0.096	2.213e-4
∑	5.050e-3	-	-

During the BER estimation the following cases were considered:

1. Ideal synchronization without equalizer for MCS0 and MCS10 (**Fig. 12**, Ideal sync.).
2. Preamble based synchronization with LS equalizer and zero residual frequency offset (Fig. 12, MCS10, RFO = 0).
3. Preamble based synchronization with LS equalizer and TF-DFL RFO compensation (Fig. 12, MCS10, TF-DFL).
4. Preamble based synchronization with LS equalizer and pilots-based RFO compensation (Fig. 12, MCS10, Pilots).
5. Preamble based synchronization with LS equalizer and pilots-based RFO compensation for traveling pilots (Fig. 12,

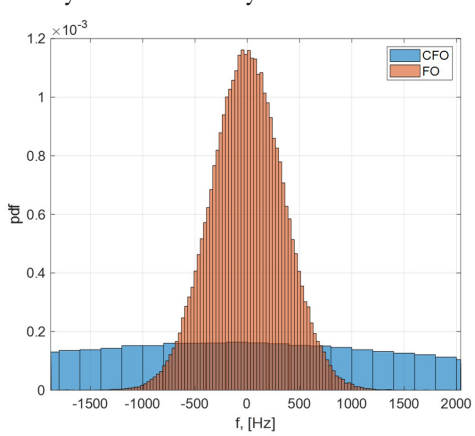


Fig. 11. *PDF of sum of coarse and fine frequency estimates.*

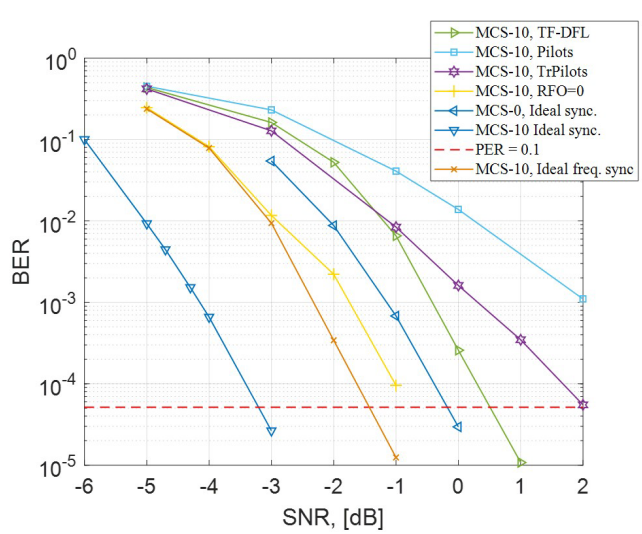


Fig. 12. *BER curves of our receiver model.*

MCS10, TrPilots)

6. Preamble based timing with LS equalizer and zero frequency offset (Fig. 12, MCS10, Ideal freq. sync.).

4. DISCUSSION

The probability of coarse timing failure that was obtained (Table 2) is close to the probability presented in the working group (TGah) [16].

The synchronization errors and LS equalizer errors lead to 2.4 dB degradation of receiver noise immunity relative to the theory.

Besides, a considerable contribution is made with timing systems failures that are illustrated with (MCS10, Ideal freq. sync.) curve in Fig. 12. In this case, degradation is around 1.8 dB.

At first, it is not clear why there is a difference in noise immunity between these two synchronization cases since for both there is no RFO. Here, the relationship between time and frequency estimates from each other plays a role. Before estimating the exact timing, the frequency offset is compensated with a coarse frequency estimate. As a result, a significant inaccuracy in the frequency estimate leads to an error in time estimate since the last one evaluated with cross-correlation.

Hence, the potential gain from improving frequency estimation systems is 0.6 dB. At the same time, the main contribution to degradation is made with fine estimation systems, since the probabilities of coarse and fine differ by about 2 orders of magnitude. As for the gain due to the improvement of timing systems, the probability of an error in the fine timing system is about an order of magnitude higher than in the coarse timing (**Table 2**), and it can be said that the improvement of the fine timing will give a gain in the receiver's noise immunity.

For the case of receiver operation with RFO compensation systems, the minimum loss (TF-DFL) relative to the theoretical boundary of MCS10 is very significant and is 3.7 dB. Even relative to the idealized case for MCS0, the TF-DFL method applied to MCS10 loses 0.7 dB.

Comparing the receiver with the TF-DFL algorithm with zero RFO case gives 1.3 dB loss to the first one. Therefore, even within the framework of the studied preamble-based synchronization systems, it is possible to increase the receiver's noise immunity by applying more advanced RFO compensation methods.

Comparing the noise immunity of classical pilot-based and TF-DFL methods, it is seen that TF-DFL shows better results, and the classical method does not help even traveling pilots (~3.5 dB gained). TF-DFL algorithm shows better results starting with BER ~1e-2, as approaching the receiver working area (PER ≤ 0.1) the gain reaches 1.5 dB.

The last thing has to be noted, the fact that LS equalizer is not the best solution in terms of receiver noise immunity, so it is possible to use better channel estimation techniques to increase it.

5. CONCLUSION

In this paper, the synchronization and tracking algorithms for IEEE 802.11ah receiver have been studied. The measurement results which were obtained during simulation in MATLAB show that usage of classical preamble-based synchronization systems with LS equalizer leads to a degradation of the BER vs SNR curve by ~2.4 dB relative to the idealized theoretical case. The main contribution to degradation is made with

equalizer and fine synchronization systems. Therefore, to increase receiver noise immunity, it is necessary to optimize these systems.

The best result in RFO compensation is achieved with TF-DFL method (PER = 10% for SNR = 0.5 dB). At the same time, other more optimal solutions for RFO compensation are possible.

REFERENCES

1. 802.11ah-2016 - *IEEE Standard for Information technology-Telecommunications and information exchange between systems - Local and metropolitan area networks--Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: Sub 1 GHz License Exempt Operation.*
2. Wang Y, Sun S, Tan PH, Kurniawan E. Baseband receiver design for IEEE 802.11ah. *TENCON 2017-2017 IEEE Region 10 Conference*, Penang, Malaysia, 2017. DOI: 10.1109/TENCON.2017.8227976.
3. Hajjar CE. *Synchronization Algorithms for OFDM Systems (IEEE802.11ah, DVB-T) Analysis, Simulation, Optimization and Implementation Aspects*. Erlangen, 2007, 166 p.
4. Kuang L, Ni Z, Lu J, Zheng J. A Time-Frequency Decision-Feedback Loop for Carrier Frequency Offset Tracking in OFDM Systems. *IEEE Transactions on wireless communications*, 2005, 4(2):367-373.
5. Jimenez VPG, Armada AG, Gonzalez-Serrano FJ. Design and Implementation of Synchronization and AGC for OFDM-based WLAN Receivers. *IEEE Transactions on Consumer Electronics,* 2004, 50(4):1016-1025.
6. Halasz D. IEEE P802.11 Wireless LANs. *Categories of TGah Use Cases and Straw Polls,* IEEE 802.11-11/0301r2, 2011.
7. Yusupov RR. Modelirovanie peredachi trafika mezhmashinnogo vzaimodeistviya v setyakh Wi-Fi HaLow s ispolzovaniem mekhanisma okna ogranichennogo dostupa v rezhime s peresecheniem granic. [M2M traffic modeling in Wi-Fi HaLow Networks with usage of restricted acces window mechanism in cross slot boundary mode]. *Proceedings of the 42nd Interdisciplinary School-Conference of IITP RAS "Information Technologies and Systems* 2018", p. 430-439 (in Russ.).
8. Loginov VA, Lyakhov A, Khorov E, Analiticheskaya model raboty mekhanizma mgnovennoy retranslyacii paketov v setyakh IEEE 802.11ah [Transmission opportunity (TXOP) analytical model in IEEE 802.11ah networks]. *Proceedings of the 39nd Interdisciplinary School-Conference of IITP RAS "Information Technologies and Systems* 2015, p. 1140-1453 (in Russ.).
9. Hazmi A, Rinne J, Valkama M. Feasibility Study of IEEE 802.11ah Radio Technology for IoT and M2M use Cases. *IEEE Globecom Workshops,* p. 1687-1692, Anaheim, CA, USA, 2012. DOI: 10.1109/GLOCOMW.2012.6477839.
10. Aust S, Prasad RV, Niemegeers IGMM, Outdoor Long-Range WLANs: A Lesson for IEEE 802.11ah. *IEEE Communications Surveys & Tutorials,* 2015, 17(3):1761–1775.
11. Porat R. Link Budget, IEEE 802.11-11/0552r2.
12. Perahia E, Stacey R. *Next Generation Wireless LANs 802.11n and 802.11ac*, 2nd ed. Cambridge, U.K., Cambridge Univ. Press, 2013.

13. Gupta RK, Jain A, Singodiya P. Bit error rate simulation using 16 qam technique in matlab, *International Journal of Multidisciplinary Research and Development*, 2015, 2(5):59-64.
14. Kleine J. *Statisticheskie metodi v imitacionnom modelirovanii [Statistical methods in simulation modeling]*. Iss. 2. Moscow, Statistika Publ., 1978, 335 p.
15. Hwang J. *Simplified Channel Estimation Techniques for OFDM Systems with Realistic Indoor Fading Channels*. Waterloo, Ontario, Canada, 2009.
16. Vermani S et al. Preamble Format for 1 MHz. *IEEE-802.11-11/1482r4*, 2012.

DOI: 10.17725/rensit.2020.12.287

Detection of DoS attacks caused by CONNECT messages of MQTT protocol

Dmitrii I. Dikii

St. Petersburg National Research University of Information Technologies, Mechanics and Optics (University ITMO), http://www.itmo.ru

St. Petersburg 197101, Russian Federation

E-mail: dimandikiy@mail.ru

Received September 25, 2019, reviewed on March 10, 2020, after finalization on March 30, 2020 accepted April 13, 2020

Abstract. **Detecting DoS attacks within the Internet of Things is an urgent task to ensure the security of this infrastructure. The malefactor, undertaking the attack, generates a large number of connection requests to the Internet of Things network based on the MQTT protocol. This makes the gateway unavailable for other users. The author discusses the approaches and methods of detecting DoS attacks within the Internet, in general, as well as within the Internet of Things, in particular. The method of feature vector generation for detecting DoS attacks based on the network traffic analysis was suggested. The feature vector consists of parameters of message transmission frequency within a time interval from a device with the same IP-address. The multilayer perceptron, the random forest algorithm, the support vector machine are classifiers in this study. The author constructed an experimental assembly to generate training and testing sets with the supplied parameters. The experiment results showed: in order to achieve maximum classification accuracy, the dimension increase of the feature vector is not required. A comparison of the mentioned above algorithms by the F1-score value was carried out, which proved the artificial neural network – the multilayer perceptron – to be the best classifier. At that, the time interval, on which the feature vector generation is based, must be higher than 1.5 seconds for the accuracy to be over 0.99 under the legal device connection frequency once per second. The research gave positive results of implementing the reviewed classifiers based on the suggested feature vector to detect DoS attacks.**

Keywords: **Internet of things, DoS, MQTT, machine learning, random forest, multilayer perceptron, support vector machine, telecommunication, attack detection**

UDC 004.052.3

Acknowledgments. The reported study was funded by RFBR, project number №19-37-90051

For citation: Dmitrii I. Dikii. Detection of DoS attacks caused by CONNECT messages of MQTT protocol. *RENSIT,* 2020, 12(2):287-296. DOI: 10.17725/rensit.2020.12.287.

Contents

1. INTRODUCTION

Recently, there has been a significant popularity increase of technologies employed in the Internet. One of those technologies is the Internet of Things [1]. The main characteristic of the technology that allows uniting a variety of projects under a single name – the Internet of Things – is a possibility for a large number of devices, functioning without an operator, to communicate to carry out a single common task. The devices mentioned below must have only the essential capabilities. This makes them significantly cheaper than common workstation computers (personal computers, smartphones, etc.). Certain devices in the Internet of Things network function on an independent power supply. This imposes limits on the employment of such devices from the energy saving point of

view. Data transfer technologies and protocols that significantly reduce energy demand of the terminal device are developed to increase its operational life from an independent power supply. Research in the field of the Internet of Things networks encloses the whole protocol stack of the OSI model [2]. One of those protocols is the MQTT (message quality telemetry transport), which was developed by the OASIS alliance [1]. Currently, the most widespread version is the MQTT protocol 3.1.1.

Along with the tendency to simplify protocols for Internet of Things devices, there is an increase in information security threats. Information circulating in the Internet of Things network remains plaintext. This may lead to negative consequences from the information owner's side. The most striking case in point is the medical field. Papers [4, 5] present arguments on security enhancement namely in healthcare. One of presented methods is to enhance device authentication.

Apart from threats common to the Internet of Things networks, threats typical to all devices connected to the Internet have to be noted. Generally, those are attacks such as a man-in-the-middle, phishing, viruses, trojans, etc. Another threat is the distributed DoS attack. Its main feature is that a large number of network devices send requests to the victim device. Due to the exceedance of the request-processing maximum per time interval, the victim does not handle the load and becomes unavailable to other devices. According to the Kaspersky Laboratory analytics, the malware based on the Mirai botnet has become the most widespread around the world by the end of first quarter of 2019 [6]. As part of the botnet behavior, the DoS attack becomes an immediate threat.

Thus, deploying the Internet of Things network infrastructure in an organization or at home, one should consider classical threats of information security as well as specific ones of the Internet of Things networks and their combinations.

In this paper, the author discusses the problem of the Internet of Things network devices' excessive use of the MQTT protocol capabilities to employ the DoS attacks.

Nowadays, the following attack classification is given:

- Bandwidth exhaustion attack;
- Victim resource exhaustion attack;
- Infrastructure attack;
- Zero day attack [7].

Malefactors most frequently use two types of DoS attacks. The first is the bandwidth saturation with information until the legitimate source signal does not reach the recipient. The second type of attack exploits the vulnerability of other protocols to exhaust the server's resources: memory space, CPU usage time. Moreover, one can implement the network and transport layer protocols as well as the application ones, such as the HTTP, for the second type of attack. A striking example is the TCP SYN attack, when the malefactor sends a request to establish a connection via the TCP protocol, but instead of specifying his own IP address, a nonexistent one is specified. The server stands by to establish a connection, but does not receive feedback overlong. However, the information concerning unestablished connection is saved on the server side, thus, leading to the victim bandwidth exhaustion.

Preventive measures to secure information systems from such attacks can be divided into two stages:

- Detection;
- Counter acting.

Detection is carried out through network traffic analysis. "Hop count" packet filtering method has gained the most popularity [8]. In this method, the number of TCP packets and statistical parameters is assessed: SYN flag, TTL, the source and destination addresses, etc. Paper

[9] reviews the DoS attack detection method and the security against it, consisting of MAC-addresses filtering and cryptographic processing.

Methods based on artificial intelligence and machine learning are more rapidly suggested to detect attacks. Thus, the authors propose to employ swarm algorithms in paper [10]. The accuracy of DoS attack detection by the suggested method is 0.75-0.80.

TCP traffic is most often analyzed by estimating the server response time for normal traffic and for attacks to protect Internet resources. During an attack, the server response time significantly increases. This fact is the basis for traffic classification. For example, using the LS-SVM algorithm made it possible to achieve classification accuracy of over 0.92 [11].

Many other papers reviewed the support-vector machines (SVM) for the DoS attacks detection. For example, the authors of [12] were able to gain 99% of positive attack identification from the TCP traffic employing the SVM on the DARPA database. The authors of [13, 14] were able to gain similar abnormal traffic detection accuracy employing the support-vector machines. The authors of paper [15] review the variations of this method. In this case, the accuracy of the presented methods is more than 0.92. The difference between the studies dedicated to employing SVM to detect DoS attacks is in the distinctive solutions to form the feature vector. Paper [16] presents a study on the influence of various features on the classification accuracy. Thus, the choice of the feature vector is the main factor affecting accuracy. The SVM algorithm or its variations showed positive results in detecting attacks under the TCP traffic analysis.

Another approach to detect DoS attacks is to employ artificial neural networks (ANN). There is a great number of artificial neural networks variations. The most widespread model, the multilayer perceptron (MLP), is reviewed in paper [11]. The comparison of the SVM and the ANN has shown that the latter has lower accuracy and requires more time to make a decision [17]. Paper [18] presents the results of the experiment on employing the ANN to detect anomalies in traffic via the TCP and ICMP protocols. The approach suggested by the authors achieved a 0.98 accuracy of detecting attacks. Moreover, the ensemble of recurrent artificial networks is used [16].

The random forest (RF) algorithm class and the decision trees are used to detect attacks. This approach shows good results. For example, the detection accuracy in paper [20] is over 0.96. Similar studies [21-23] present high detection accuracy. Fuzzy logic approaches are also used to detect DoS attacks [24].

Methods of abnormal traffic detection in the Internet of Things networks are based on data analysis of transport and network protocols, as described in [25, 26]. However, the Internet of Things networks use other layer protocols that are vulnerable to DoS attacks. Such as application layer protocols (CoAP, MQTT), and protocols of lower layers (for example, LoRa). Attacks on the physical and data-link layers are most widespread in the wireless sensor networks. For example, an attack aimed at resource exhaustion is described in [27]. Another attack common to the Internet of Things network is "blackhole". In this case, a device communicates to other devices in the network that its node has the shortest route to deliver a packet. However, all packets delivered to this node will be dropped [28]. In addition, there are attacks that create jamming for information transmission via radio channels, thus, causing DoS [29].

2. METHODS AND MATERIALS

In relation to the application layer protocol MQTT, one can note its prepossession to DoS attacks. Generally, this is employed by increasing the load on the network elements to disrupt the communication between devices. The protocol functions on the "publish-subscribe" pattern. Thus, the network has a key element called

the gateway. It is responsible for redirecting messages from the sender to the recipient. As all messages pass the gateway it is the most vulnerable element to the attack. Research was carried out concerning the influence of message parameters (flags, message amount, etc.) on gateway sustainability under high loads. Authors in most studies reviewed only messages of PUBLISH type with the following parameters:

- Quality of service (QoS) [30, 31];
- Number of subscribers [32];
- Message payload size [33, 34];
- Cryptographic processing of messages [32].

On the contrary, the process of device connection to the gateway is not taken into account. When simultaneously sending a large amount of connection requests (CONNECT messages), the gateway may not be able to handle the load. As a result, legal devices will not be able to connect to the gateway to send or receive messages [35]. Thus, in the networks that operate on the MQTT protocol one must detect abnormal device behavior on all stages of protocol operation.

The purpose of this work is to develop a method of detecting a denial-of-service attack caused by abnormal behavior of network devices by exploiting CONNECT messages of the MQTT Protocol using machine learning algorithms. To achieve this purpose, first, the task of choosing the optimal feature vector is solved. The second task to be solved is to determine the most effective classification method. The following algorithms were considered as classifiers in this study: multilayer perceptron, random forest, and support vector machine with a radial basis function of the kernel, programmatically implemented on the basis of the WEKA project [36]. To generate training and testing data sets, an experimental assembly was created (**Fig. 1**), consisting of a gateway, communication equipment, and several computers that simulate the behavior of many Internet of things devices using the Paho-mqtt

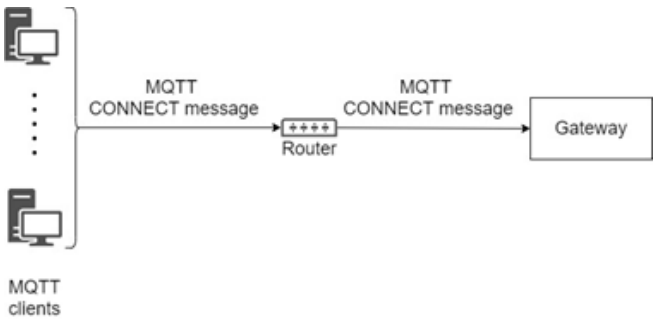


Fig. 1. *Experimental assembly scheme.*

framework [37]. The gateway was a Raspberry Pi 3 model B microcomputer with Moquette project software on JAVA language [38].

A feature vector must be formed, for the message to be correctly classified. As the malefactor merely needs the gateway address and the port number to send a connection request, thus, service information as device ID and username can be generated automatically and are not considered. Therefore, the main parameters describing CONNECT message via the MQTT protocol are:

- Sender address as the IP address. It is required to keep a block list of addresses from which the attack originates. In this case, the IP address is used as a tag;
- Number of connection requests per a time interval;
- Connection-time mean between connection requests per a time interval;
- Binary value determining the employment of cryptographic processing via TLS protocol, which significantly affects time of connecting to the gateway: 0 – TLS protocol is not employed, 1 – TLS protocol is employed.

Therefore, the feature vector of the connection message consists of three main parameters per a single analyzed time interval (hereafter m).

The choice of a time interval plays an important role in forming a feature vector. Moreover, it can be not a single time interval, but a complex. Thus, the feature vector dimension can be increased and be estimated by the formula:

$$W = 1 + 3k, \quad (1)$$

where k – the number of time intervals m.

A time interval m is a period between the moment when the gateway received a message and the moment of the predetermined number of milliseconds to this instance. An example of forming a body of three time intervals is depicted on **Fig. 2**.

Legal traffic was generated based on device behavior simulation of a real network with consideration to installed secure and insecure connections via the TLS protocol. The simulation of legal traffic depends on practical application conditions of the network, thus, the training data set will vary depending on the estimated maximum network load. The abnormal traffic was simulated by generating a high message flow with requests to connect via an insecure channel as well as a secure one by the TLS protocol. Channel security selection is determined randomly for each connection.

A test data set consisting of legal and abnormal traffic examples was collected to determine the best classifier. A network operation scenario in a standard mode consisting of ten thousand connections was used to collect the data set. In addition, there was a scenario with a potential attack consisting of five thousand sequentially sent CONNECT messages.

The question of selecting time intervals m, based on which the feature vector will be formed remains. The following methodology for selecting optimal time intervals was suggested.

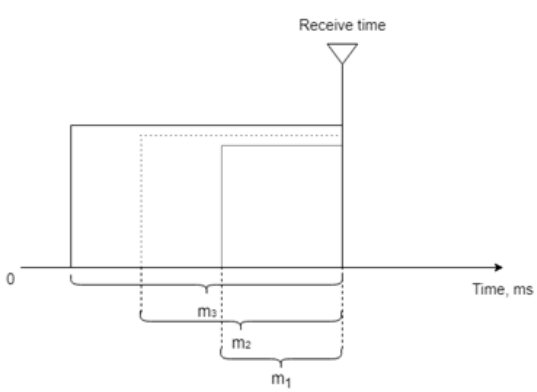


Fig. 2. *Time interval definition scheme for feature vector,* $m_3 > m_2 > m_1$.

As a first step, expert analysis determines a finite set of time intervals M with natural values ($M \in N$).

On the second step, three classifiers (MLP, RF, SVM) are trained for each value of $m \in M$ and those classifiers undergo proving on the test data set.

In the third step, the classification quality is evaluated by calculating the F1-score, which is a weighted average of precision and recall. This metric is widely used in evaluating the quality of binary classification for machine learning methods, as is shown in [24, 39]. To calculate this value, the classification results of the number of correctly and incorrectly classified messages on the test dataset are used (**Table 1**, where TP – the number of legitimate messages recognized correctly; TN – the number of attack messages recognized correctly; FP – the number of abnormal messages recognized incorrectly; FN-the number of legal messages recognized incorrectly).

Then classification precision is calculated by the formula:

$$\text{Precision}=TP/(FP+TP) \quad (2)$$

and recall by the formula:

$$\text{Recall}=TP/(TP+FN) \quad (3)$$

Knowing those values, one can calculate the F1-score using the formula:

$$F=(2\times\text{Precision}\times\text{Recall})/(\text{Precision}+\text{Recall}) \quad (4)$$

On the fourth step, a finite set S is formed with unique combinations of nonrecurring elements $m \in M$ of the length l, such that $2 \leq l \leq |M|$.

Table 1

Confusion matrix

	Legal messages	Abnormal messages
Correctly classified messages	TP	TN
Incorrectly classified messages	FN	FP

On the fifth step, steps 2 and 3 are repeated with the determined time intervals of the set *S*. The best time interval combination, under which the highest F1-score value is reached.

3. RESEARCH RESULTS

The following initial time interval set for *M* was established for the simulated network {20, 50, 100, 150, 200, 250, 500, 1000, 1500, 2000, 3000} ms.

The training data set consists of two arrays. The first one includes information about legal message flow of the connection to the gateway. Whereas the second has information about a flow similar to the DoS attack. The following model was established to generate a legal data flow. A time interval *I*, during which at least a single connection message is sent guaranteed, is established. The dispatch time *i* is randomly determined (via equal probability distribution so that the data set contains examples with both small and close to maximum values of the delay between messages) from the interval *I* ($i \in I$). Therefore, the time difference between the dispatches of two sequential messages is defined as follows:

$$\Delta T = i + t, \quad (5)$$

where *t* – the time required to process the messages and receive a response from the gateway (via an insecure channel not over 50 ms in average, via a secure one – not over 700 ms for the considered experimental assembly), *i* – random time delay, not over the maximum value of *I* interval.

Two intervals were considered, $I = [0, 1000]$ ms and $I = [0, 500]$ ms, to determine the impact of a training sample containing legitimate traffic on the quality of classification,.

The author simulated a situation where a large CONNECT message flow was sent to the gateway over a short period to generate the second data array containing data about abnormal data flow. The training data set

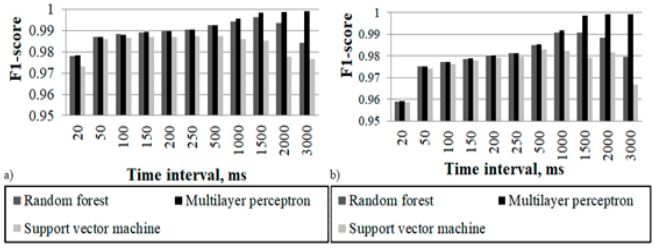


Fig. 3. *Classification results of legal messages generated from the interval a)* $I = [0, 1000]$ *ms, b)* $I = [0, 500]$ *ms.*

consists of ten thousand messages for legal data flow and five thousand messages for simulating the DoS attack. The classification results by the testing data set under one time interval $m \in M$ is presented on **Fig. 3**.

The results of the conducted experiment show the classifier based on the multilayer perceptron to be the best one. An increase of the time interval, during which traffic statistics is collected, leads to the increase of the F1-score value. For example, the value reaches 0.9989 ± 0.0001 under the interval of two seconds.

There is a complex dynamic to the random forest algorithm. The F1-score increases as the time interval increases from 20 ms to 1500 ms. However, further increase of the time interval to three seconds leads to the classifiers characteristics degradation. The F1-score value maximum is achieved at interval m = 1500 ms and is equal to over 0.9934 ± 0.0027.

The support-vector machine was the worst algorithm to cope with classification. The F1-score value was lower than that of other algorithms under every considered time interval. The F1-score value maximum is achieved at m = 500 ms and the classification accuracy equals 0.985 ± 0.0021. The classification accuracy decreases, when increasing the time interval to three seconds.

The employment of a set of analyzed intervals $m \in M$ did not bring a significant positive effect. **Figure 4** presents F1-score values of the considered algorithms under the following interval sets: {200, 250, 500},

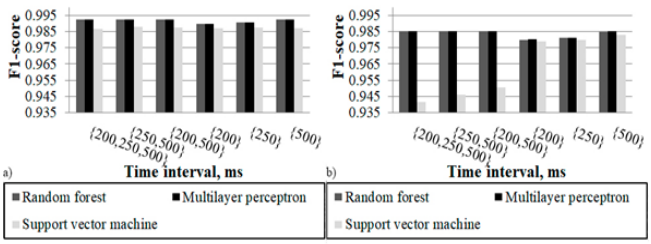


Fig. 4. *Classification results for legal messages generated from the interval a) I = [0, 500] ms, b) I = [0, 1000] ms.*

{250, 500}, {200,500}. Intervals {200}, {250}, {500} are shown for comparison purposes. Thus, the employment of a set of intervals does not increase classification quality, but often decreases it. For example, when employing the support-vector machine under the legal request frequency of the I = [0, 1000] interval, one can observe evident negative dynamics of the F1-score in cases of feature space increase. The F1-score values are lower or insignificantly higher than the values under classification by a single widest interval in most other cases.

Therefore, the use of a set of intervals to form a feature vector of a larger dimension is inefficient.

4. DISCUSSION

In the scope of the conducted study, the detection methods of DoS attacks employed in the Internet networks as well as the Internet of Things networks were analyzed. The suggested feature vector for CONNECT messages by the MQTT protocol consists of three main parameters: the amount of messages per time interval, mean connection-time between two sequential messages, the mean value of the parameter responsible for the TLS protocol employment when creating a cryptographically secure channel per a time interval, and a tag-parameter – sender's IP address.

The following algorithms were considered as classifiers: multilayer perceptron, random forest algorithm, support-vector machine. The experiment based on generated training and testing data sets showed that all algorithms cope with the traffic classification task with accuracy over 0.90. The multilayer perceptron model had the best classification quality. The F1-score values increased according to the time interval increase, during which traffic statistics was collected and the feature vector was formed. The random forest algorithm coped with the classification with worse values. The time interval increase up to 1.5 seconds has a positive effect on the dynamics F1-score. However, further increase of the time interval leads to the F1-score value decrease. The worst to cope with the classification was the support-vector machine. The dynamics of F1-score is similar to the one of random forest algorithm. The F1-score maximum is gained at the 500 ms interval. The employment of feature vectors of larger dimension is inefficient as the classification characteristics may not only remain the same but also degrade.

5. CONCLUSION

Therefore, among all considered approaches and algorithms of detecting DoS attacks by the suggested feature vector (the feature vector in this case will have dimension 4) it is recommended to employ the multilayer perceptron. That model showed the best results in contrast of other considered methods. However, the classification quality increases with the time interval during which traffic statistics is collected. However, it is worth noting that increasing this interval will lead to a large computational and time-consuming cost of training the model and making a decision. The quality of classification based on the random forest algorithm or the support vector machine with the radial basis function of the core is worse than that of the multilayer perceptron, but the values of the F1-score are high enough to be used.

Further research will be dedicated to studying DoS attacks caused by the abuse of other types of MQTT protocol messages.

REFERENCES

1. Ashton K. That 'Internet of Things' Thing. *RFID Journal*, 2009, 22:97–114.
2. Standart "ISO/IEC 7498-1:1994 [ISO/IEC 7498-1:1994] Information technology — Open Systems Interconnection — Basic Reference Model: The Basic Model". *ISO/IEC Information Technology Task Force (ITTF) web site*, 1994.
3. Standart ISO/IEC 20922:2016 Information technology – Message Queuing Telemetry Transport (MQTT) v3.1.1. *ISO/IEC Information Technology Task Force (ITTF) web site,* 2016.
4. Albalawi U, Joshi S. Secure and Trusted Telemedicine in Internet of Things IoT. *Proceedings of 2018 IEEE 4th World Forum on Internet of Things* (WF-IoT), 2018, pp. 30-34. DOI: 10.1109/WFIoT.2018.8355206.
5. Wazid M, Kumar Das A, Khurram Khan M, Al Dhawailie AlGhaiheb A, Kumar N, Vasilakos AV. Secure Authentication Scheme for Medicine Anti-Counterfeiting System in IoT Environment. *IEEE Internet of Things Journal,* 2017, 4(5):1634-1646. DOI: 10.1109/JIOT.2017.2706752.
6. Chebyshev V, Sinitsyn F, Parinov D, Larin B, Kupreev O, Lopatin E. Development of information threats in the first quarter of 2019. *Statistics. Kaspersky Security Bulletin 2019. Statistics*. Threat reports URL: https://securelist.ru/it-threat-evolution-q1-2019-statistics/94021/ (дата обращения: 20.08.2019).
7. Mahjabin T, Xiao Y, Sun G, Jiang W. A survey of distributed denial-of-service attack, prevention, and mitigation techniques. *International journal of distributed senssor networks*, 2017, 13(12):1-32. DOI: 10.1177/1550147717741463.
8. Jin C., Wang H., Shin K.G Hop-Count Filtering: An effective defense against spoofed DDoS traffic. *Proc. of the ACM Conf. on Computer and Communications Security,* 2003, pp. 30-41. DOI: 10.1145/948109.948116.
9. Prakash A, Satish M, Sri Sai Bhargav T, Bhalaji N. Detection and Mitigation of Denial of Service Attacks Using Stratified Architecture. *Proc. of the 4th Intern. Conf. on Recent Trends in Computer Science & Engineering Detection Procedia Computer Science,* 2016, 87:275-280. DOI: 10.1016/J.PROCS.2016.05.161.
10. Sharma S, Gupta A, Agrawal S. An Intrusion Detection System for Detecting Denial-of-Service Attack in Cloud Using Artificial Bee Colony. *Proc. of the Intern. Congress on Information and Communication Technology, Advances in Intelligent Systems and Computing,* 2016, pp. 137-145. DOI: 10.1007/978-981-10-0767-5_16.
11. Sahi A, Lai D, LI Y, Diykh M. An Efficient DDoS TCP Flood Attack Detection and Prevention System in a Cloud Environment. *IEEE Access,* 2017, 5:6036-6048. DOI: 10.1109/ACCESS.2017.2688460.
12. Mukkamala S, Sung AH. Detecting Denial of Service Attacks Using Support Vector Machines. *Proc. of the 12th IEEE Intern. Conf. on Fuzzy Systems,* 2003, pp.1231-1236. DOI: 10.1109/FUZZ.2003.1206607.
13. Manuel S. Hoyos Ll, Gustavo AIE, Jairo IV, Castillo OL. Distributed Denial of Service (DDoS) Attacks Detection Using Machine Learning Prototype. *Proc. of the 13th Intern.Conf., Advances in Intelligent Systems and Computing,* 2016, pp. 33-41. DOI: 10.1007/978-3-319-40162-1_4.
14. Kim D, Lee KY. Detection of DDoS Attack on the Client Side Using Support Vector Machine. *Intern. J. of Applied Engineering Research,* 2017, 12(20):pp. 9909-9913.
15. Xu X, Wei D, Zhang Y. Improved Detection Approach for Distributed Denial of Service Attack Based on SVM. *Proc. of the 3th Pacific-Asia Conference on Circuits, Communications and System (PACCS)*, 2011, pp. 1-3. DOI: 10.1109/PACCS.2011.5990284.
16. Chan APF, Ng WWY, Yeung DS, Tsang ECC. Refinement of rule-based intrusion detection system for denial of service attacks by support vector machine. *Proc. of the 13rd*

Intern. Conf. on Machine Learning and Cybernetics, 2004, pp. 4252- 4256. DOI: 10.1109/ICMLC.2004.1384585.

17. Tsang GCY; Chan PPK; Yeung DS; Tsang ECC. Denial of service detection by support vector machines and radial-basis function neural network. *Proc. of Intern. Conf. on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*, 2004, pp. 4263-4267. DOI: 0.1109/ICMLC.2004.1384587.
18. Saied A, Overill RE, Radzik T. Detection of known and unknown DDoS attacks using Artificial Neural Networks. *Neurocomputing,* 2016, 172:385-393. DOI: 10.1016/j.neucom.2015.04.101.
19. AI Islam ABMA, Sabrina T. Detection of various Denial of Service and Distributed Denial of Service Attacks using RNN Ensemble. *Proc. of 12th Intern. Conf. on Computer and Information Technology (ICCIT 2009),* 2009, pp. 603-608. DOI: 10.1109/ICCIT.2009.5407308.
20. Lakshminarasimman S; Ruswin S; Sundarakantham K. Detecting DDoS attacks using decision tree algorithm. *Proc. of 4th Intern. Conf. on Signal Processing, Communication and Networking (ICSCN)*, 2017, pp.1-6. DOI: 10.1109/ICSCN.2017.8085703.
21. Chen L, Zhang Y, Zhao Q, Geng G, Yan Z. Detection of DNS DDoS Attacks with Random Forest Algorithm on Spark. *Proc. of 2nd Intern. Workshop on Big Data and Networks Technologies Procedia Computer Science,* 2018, 134:310-315. DOI: 10.1016/j.procs.2018.07.177.
22. Idhammad M, Afdel K, Belouch M. Detection System of HTTP DDoS Attacks in a Cloud Environment Based on Information Theoretic Entropy and Random Forest. *Security and Communication Networks*, 2018, 2018:1-13. DOI: 10.1155/2018/1263123.
23. Cheng J, Li M, Tang X, Sheng VS, Liu Y, Guo W. Flow Correlation Degree Optimization Driven Random Forest for Detecting DDoS Attacks in Cloud Computing. *Security and Communication Networks,* 2018, 2018:1-14. DOI: 10.1155/2018/6459326.
24. Haripriya AP, Kulothungan K. Secure-MQTT: an efficient fuzzy logic-based approach to detect DoS attack in MQTT protocol for internet of things. *Journal on Wireless Communications and Networking*, 2019, Vol. 90. DOI: 10.1186/s13638-019-1402-8.
25. Doshi R, Apthorpe N, Feamster N. Machine Learning DDoS Detection for Consumer Internet of Things Devices. *Proc. of IEEE Symposium on Security and Privacy Workshops,* 2018, pp. 29-35. DOI 10.1109/SPW.2018.00013.
26. Meidan Y, Bohadana M, Mathov Y, Mirsky Y, Breitenbacher D, Shabtai A, Elovici Y. N-BaIoT: Network-based Detection of IoT Botnet Attacks Using Deep Autoencoders. *IEEE Pervasive computing,* 2018, 13(9):1-8. DOI: 10.1109/MPRV.2018.03367731.
27. Mallikarjunan KN, Muthupriya K, Shalinie SM. A survey of Distributed Denial of Service attack. *Proc. of 10th Intern. Conf. on Intelligent Systems and Control (ISCO)*, 2016, pp. 1-6. DOI: 10.1109/ISCO.2016.7727096.
28. Cetinkaya A, Ishii H, Hayakawa T. An Overview on Denial-of-Service Attacks in Control Systems: Attack Models and Security Analyses. *Entropy*, 2019, 21:1-29. DOI:10.3390/E2102021029.
29. Wood AD, Stankovic JA. Denial of Service in Sensor Networks. *Computer,* 2002, 35(10):54-62. DOI: 10.1109/MC.2002.1039518.
30. Chifor B, Patriciu V. Mitigating DoS attacks in publish-subscribe IoT networks. *Proc. of Conf.: Electronics, Computers and Artificial Intelligence,* 2017, pp. 1-6. DOI: 10.1109/ECAI.2017.8166463.
31. Handosa M, Gracanin D. Performance evaluation of mqtt-based internet of things system. *Proc. of Winter Simulation Conference,* 2017, pp. 4544-4545. DOI: 10.1109/WSC.2017.8248196.
32. Fehrenbach P. Messaging Queues in the IoT Under Pressure-Stress Testing the Mosquitto MQTT

Broker. *Fakultät Informatik. Hochschule Furtwangen University*, 2017. URL: https://blog.it-securityguard.com/wp-content/uploads/2017/10/IOT_Mosquitto_Pfehrenbach.pdf.

33. Firdous SN, Baig Z, Valli C, Ibrahim A. Modelling and Evaluation of Malicious Attacks against the IoT MQTT Protocol. *Proc. IEEE Intern.Conf. on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2017, pp. 748-755. DOI: 10.1109/ITHINGS-GREENCOM-CPSCOM-SMARTDATA.2017.115.
34. Bao C, Guan X, Sheng Q, Zheng K, Huang X. A Tool for Denial of Service Attack Testing in IoT. *Proc. 8th Intern. Conf. on Information Technology in Medicine and Education (ITME)*, 2016, pp. 1-6.
35. Official web-site WEKA project. URL: https://www.cs.waikato.ac.nz/ml/weka/32 (date of the application: 11.03.2020).
36. Dikiii D.I. MQTT protocol analysis for denial of service attacks. *Scientific and technical. information bulletin technology, mechanics and optics of ITMO*. 2020, 2(2). DOI: 10.17586/2226.1494.2020.20.2.
37. Official web-site of client paho-MQTTt. URL: https://pypi.org/project/paho-MQTT/1.3.0/ (date of the application: 20.08.2019).
38. Brokers official website Moquette. URL: https://projects.eclipse.org/projects/iot.moquette (date of the application 20.08.2019).
39. Hasan M, Islam M, Zarif I, Hashem M. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. *Internet of Things,* 2019, 7:1–14. DOI: 10.1016/J.IOT.2019.100059.

DOI: 10.17725/rensit.2020.12.297

Automated Attendance Machine Using Face Detection and Recognition System

Muhanned AL-Rawi
Ibb University, https://www.ibbuniv.edu.ye/
Ibb, Yemen
E-mail: muhrawi@yahoo.com
Received February 27, 2020, reviewed March 02, 2020, accepted March 23, 2020

Abstract. **This paper serves to automate the prevalent traditional tedious and time wasting methods of marking student attendance in classrooms. The use of automatic attendance through face detection and recognition increases the effectiveness of attendance monitoring and management. This method could also be extended for use in examination halls to curb cases of impersonation as the system will be able to single out the imposters who won't have been captured during the enrollment process. Applications of face recognition are widely spreading in areas such as criminal identification, security systems, image and film processing. The system could also find applications in all authorized access facilities.**

Keywords: **automated attendance machine; face; detection and recognition**

UDC 004.931

For citation: Muhanned AL-Rawi. Automated Attendance Machine Using Face Detection and Recognition System. *RENSIT*, 2020, 12(2)297-304. DOI: 10.17725/rensit.2020.12.297.

Contents

1. INTRODUCTION

Maintaining attendance is very important in all learning institutes for checking the performance of students. In most learning institutions, student attendances are manually taken by the use of attendance sheets issued by the department heads as part of regulation. The students sign in these sheets which are then filled or manually logged in to a computer for future analysis. This method is tedious, time consuming and inaccurate as some students often sign for their absent colleagues. This method also makes it difficult to track the attendance of individual students in a large classroom environment [1,2]. In this paper, we propose the design and use of a face detection and recognition system to automatically detect students attending a lecture in a classroom and mark their attendance by recognizing their faces.

While other biometric methods of identification (such as iris scans or fingerprints) can be more accurate, students usually have to queue for long at the time they enter the classroom [2,3]. Face recognition is chosen owing to its non-intrusive nature and familiarity as people primarily recognize other people based on their facial features. This (facial) biometric system consists of an enrollment process in which the unique features of a persons' face is stored in a

database and then the processes of identification and verification. In these, the detected face in an image (obtained from the camera) is compared with the previously stored faces captured at the time of enrollment [4,5,6].

In this paper, we are setting up to design a system comprising of two modules. The first module (face detector) is a mobile component, which is basically a camera application that captures student faces and stores them in a file using computer vision face detection algorithms and face extraction techniques. The second module is a desktop application that does face recognition of the captured images (faces) in the file, marks the students register and then stores the results in a database for future analysis.

2. METHODOLOGY AND DESIGN

2.1. System design

In this design, several related components in terms of functionality are grouped to form subsystems which then combine to make up the whole system. Breaking the system down to components and subsystems informs the logical design of the class attendance system.

2.2. General overview

The flow diagram of **Fig. 1** depicts the systems operation. From Fig.1, it can be observed that most of the components utilized are similar;(the image acquisition component for browsing for input images, the face detector and the faces database for storing the face label pairs) only that they are employed at the different stages of the face recognition process.

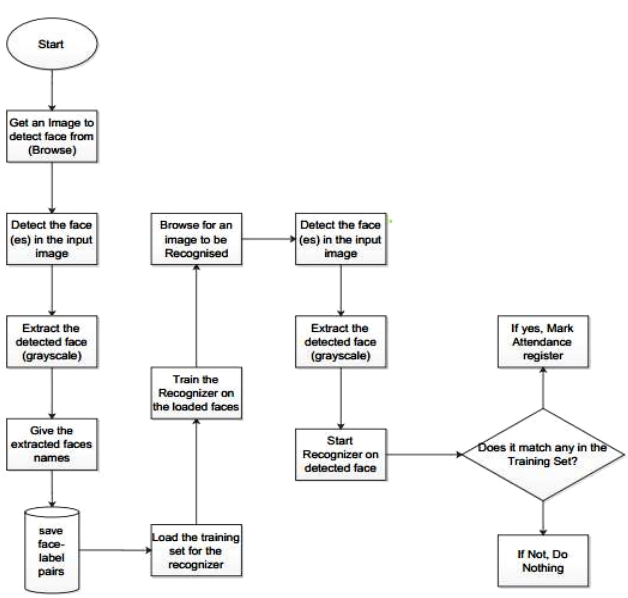


Fig. 1. *Sequence of events in the class attendance system.*

2.3. Training set manager subsystem

The logical design of the training set management subsystem is going to consist of an image acquisition component, a face detection component and a training set management component. Together, these components interact with the faces database in order to manage the training set. These are going to be implemented in a windows application form.

2.4. Face recognizer subsystem

The logical design of the face recognizer consists of the image acquisition component, face recognizer and face detection component all working with the faces database. In this, the image acquisition, and face detection component are the same as those in the training set manager sub system as the functionality is the same. The only difference is the face recognizer component and its user interface controls. This will load the training set again so that it trains the recognizer on the faces added and show the calculated Eigen faces and average face. It should then show the recognized face in a picture box.

2.5. Full mobile module logical design

This android application module which is shown in **Fig. 2** consists of a camera component, android face detector component and a SQLite database component to store the detected images. The android face detector and camera components work to detect a face from the camera input image. The image is then captured and saved in the SQLite database. This is retrieved by the image acquisition component of the desktop module.

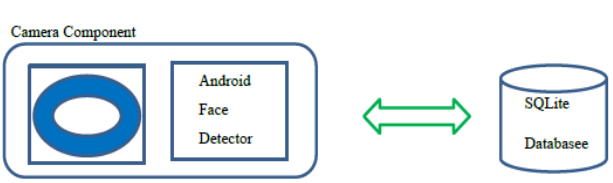


Fig. 2. *Logical design of the mobile module.*

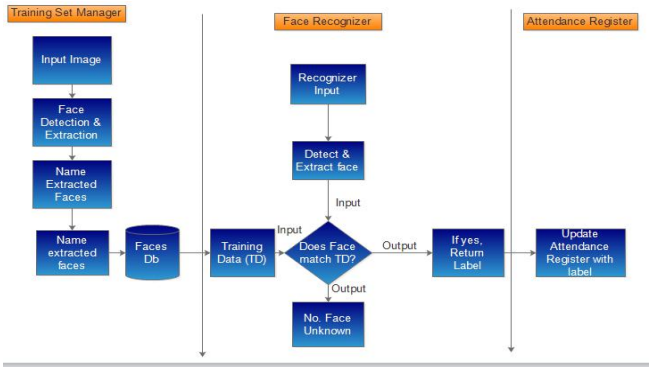


Fig. 3. *The logical design of the desktop module subsystems.*

2.6. System Architecture

Fig. 3 below shows the logical design and implementation of the three desktop subsystems

2.7 Functions of the two subsystems

The functionalities of the components are depicted in the block diagrams of **Fig. 4**. The face recognizer system consists of two major components i.e. the training set manager and the face recognizer. These two components share the faces database, the image acquisition and the face detector components; as they are common in their functionality.

We therefore partition the system in to two subsystems and have their detailed logical designs to be implemented.

2.8. Full systems logical design

The logical design of the whole system is shown in **Fig. 5.**

2.9. Tools

These tools are used in the implementation of the designed system. They've been divided in to two categories; Mobile and Desktop tools.

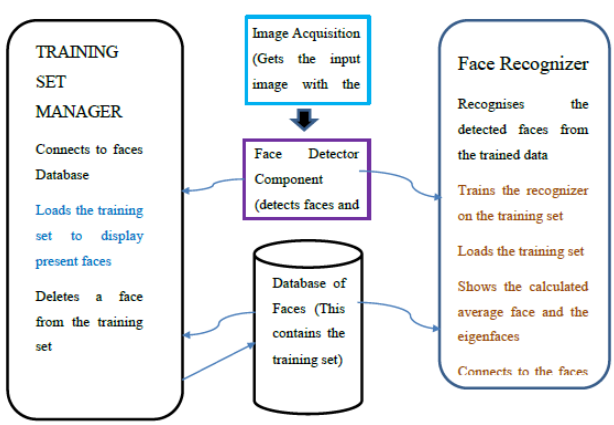


Fig. 4. *Block diagram showing functions of the components.*

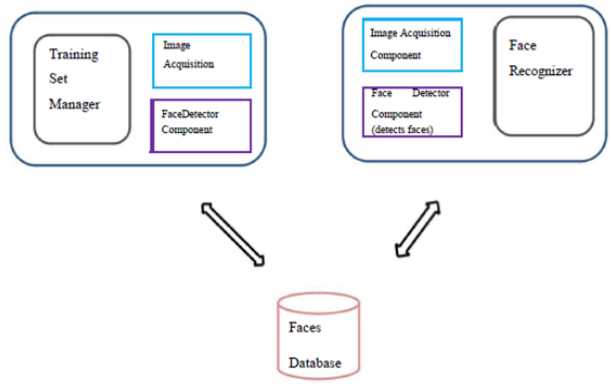


Fig. 5. *Logical design of the whole system.*

The mobile tools are the components that aid in the implementation of the mobile module. This module is responsible for capturing the students' images in a classroom environment and then storing them for further processing by the desktop module. The desktop tools are components; hardware or software that are utilized in the actual development of the desktop module. The desktop module also connects to the class attendance register which is implemented as a database management system.

3. RESULTS AND ANALYSIS

3.1. User interface of the system

3.1.1. Faces database editor

The faces database editor adds faces in the training set. The image is acquired from the highlighted box number 1 as shown in **Fig. 6** and displayed as is on step 2 on a picture box. The Regions of Interest (ROI) i.e. faces in the image is then automatically detected by drawing a light green rectangular box. In step 3, we give the extracted grayscale face from the image a face label and then add them to the training

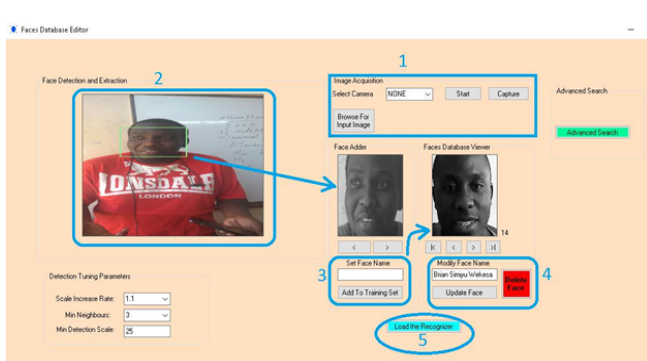


Fig. 6. *The training set editor.*

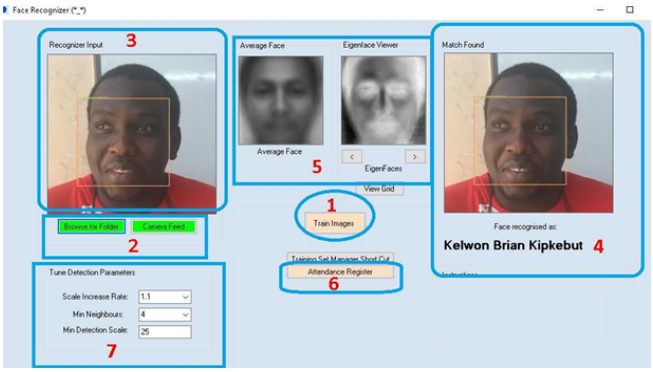


Fig. 7. *The face recognizer.*

set. In step 4, we can then modify the face label pairs in the event they are wrongly captured or even delete the faces if they are not as per the standards. Finally step 5 prepares us for the recognition stage.

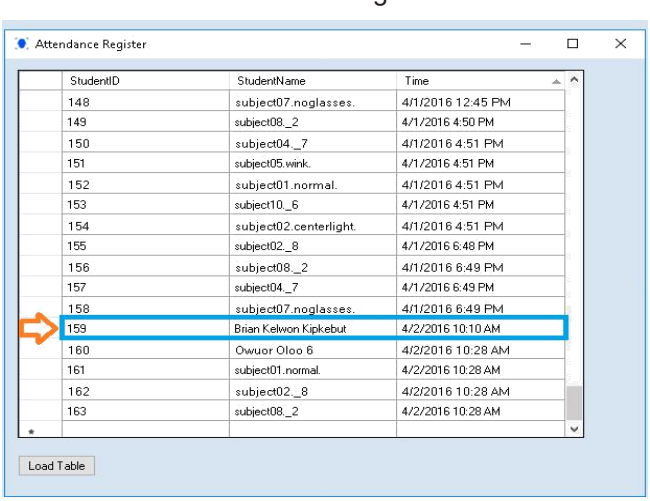
3.1.2. The face recognizer

The face recognizer compares the input face in the image captured with the faces captured during enrollment. If it is a match, it then retrieves the name associated with the input face.

Step 1 is to train the recognizer to be able to identify a face as either known or unknown. Step 2 selects the source of the image with the face to be recognized. This could be from a live camera feed or a folder with captured images. The input image with the face is then displayed in the recognizer picture box 3 as shown in **Fig. 7**. The name of the input face in the image is then displayed as shown in Step 4. The returned name of the input face, date and time are then utilized in populating the records in the attendance register database. Clicking the button of step 6 displays the register as shown in **Table 1**. The highlighted step 5 displays the computed average and Eigen faces. The arrows are used to navigate through the Eigen faces. The "View Grid" button displays the Eigen faces/vectors that had been computed from the covariance matrix in a grid form.

Table 1

Attendance register.



Selecting camera feed as the source of the input image pops up the window of **Fig. 8**. The images in the video feed are automatically detected, tracked and recognized. Images can also be added to the database from the live camera feed.

From Fig. 8, the highlighted box 1 shows the current camera view/scene. The faces and eyes in the images are automatically detected as indicated by the rectangular boxes around them. The detected face is extracted and compared with those in the database. Upon a successful match, the name associated with the face is then displayed on the upper edge of the rectangular box. The number of faces in the scene as well as their corresponding names are also shown on the highlighted box number 2. The face adder box 3 can also be used to add faces to the database.

3.2. Face detection

For group photos, a minimum neighbors' detection tuning parameter of 3 yields the best overall performance as indicated in **Fig. 9** where the physical count is 53.

The face marked by a red hexagon is not detected in the minimum neighbors' setting of 4. This is because the face is not fully displayed. Four is the highest setting which strictly returns

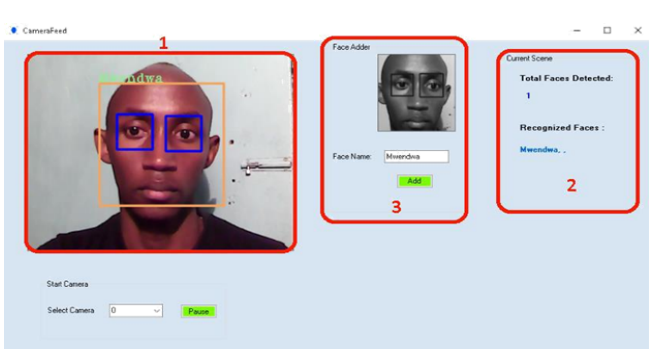


Fig. 8. *The live camera feed window.*

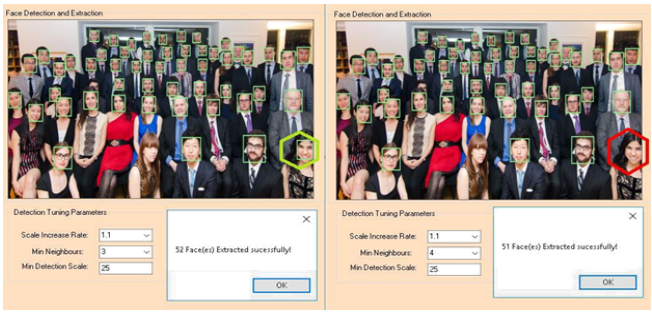
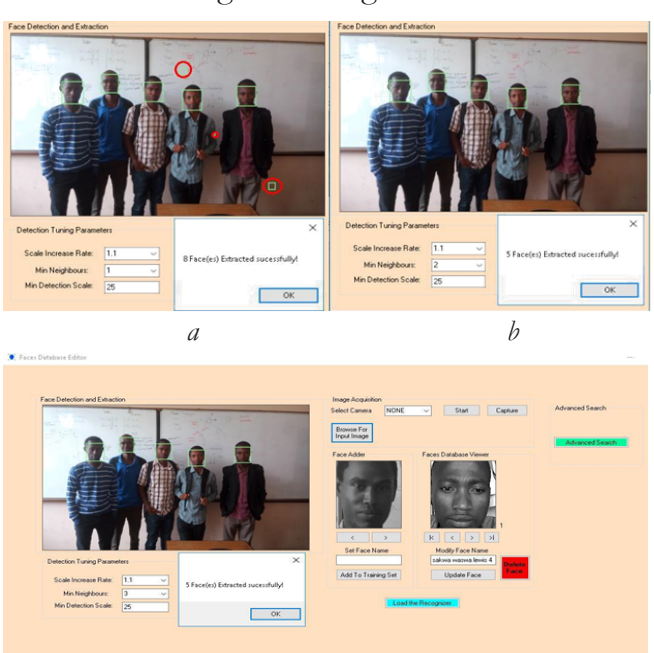


Fig. 9. *Comparison between minimum neighbors setting of three and four.*

frontal images. The second lady on the first row is not detected in either of the settings because her face is skewed to the right. The face detector only works with frontal images. 52 out of 53 images are successfully detected.

Fig. 10 shows a group photo with a minimum neighbors setting of 1 and 2. Tuning the minimum neighbors setting to 1 returns the number of faces in the images as 8; different from the physical count of 5. This is because the detector returned the slightest resmblance to a face as an actual face and hence the three face detections marked in red circles as shown in Fig. 10. Using the same image from class and incrementing the setting to 2 returned the number of detected faces as 5 which corresponded with the physical count. Increasing the setting further to 4 reduced



a *b*

c

Fig. 10. *Minimum neighbors setting of 1, 2 and 3 respectively on an image from class.*

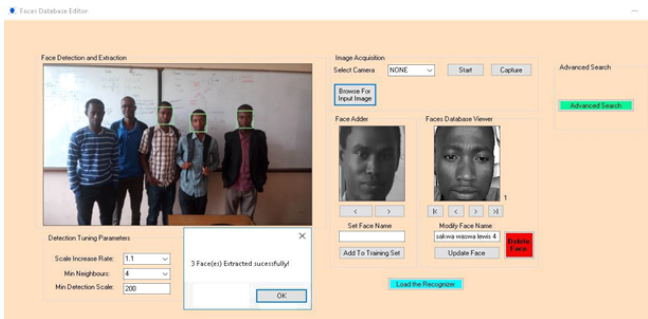


Fig. 11. *Minimum detection scale of 200.*

the number of detected faces to three.

A minimum detection scale of 25 had the best overall performance for very large group photos in terms of speed. Increasing the scale to 200 as shown in **Fig. 11** tremendously reduces the time taken to return the number of faces in an image. The minimum detection scale also makes it possible to be able to detect and recognize faces over longer and shorter distances of recognition by decreasing and increasing the scale respectively. Low detection scales waste central processing unit cycles if the size of the faces in the image is large.

The system had 100% face detection rate for different frontal faces; local as well as faces from standard faces databases like the Yale faces. The system is also able to detect bearded faces as well as faces with glasses.

3.3. Face recognition

In order to improve the recognition efficiency of the system, nine photos for each person from the standard Yale faces database are chosen for training, the remaining two photos are chosen for the testing set. Out of the fifteen subjects from the Yale faces database, twelve faces are correctly recognized. This is proportional to 80% accuracy. The faces of **Fig. 12** are not properly recognized.

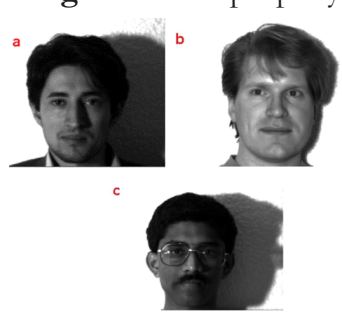


Fig. 12. *The Yale database faces that were not properly recognized.*

Fig. 13. *Images from class.*

Out of the fourteen faces of **Fig. 13**, ten are successfully recognized corresponding to a recognition accuracy of about 71.43%. The main cause of false recognition is the strength of the trained data and the illumination of the image. Face recognition is a form of machine learning and thus the larger and diverse the faces in the training set, the stronger the trained data used in recognizing faces.

Having several diverse faces of the same person with different facial expressions possible at the time of recognition creates strong training data and increases the accuracy of recognition. The lighting conditions present at the time of capturing the image to be recognized also affects the recognition results as is the case in Fig. 12 (a) and (c). Two closely identical people could also be recognized as one person unless the training data is strong.

Out of 60 faces in the database, tests are done for several subsets exclusive of the Yale dataset. The results obtained are tabulated as in **Table 2.** The percentage recognition rate is computed as the average of the percentages for the different subsets. Faces with or without glasses had no effect on the recognition rates. The mean percentage recognition rate is obtained to be 80.22%.

Table 2

Recognition results for various datasets

Dataset	No of Face	Successfully detected Faces	Successfully Recognized Faces	%Correct recognition
Center light	10	10	9	90
Left light	15	15	11	73.3
Right light	15	15	12	80
Veiled Faces	10	10	7	70
Bearded Faces	10	10	8	80
Unveiled A	20	20	17	85
Unveiled B	30	30	25	83.3

Center light faces had the best overall recognition rate at 90%. The primary issues facing most of the face detection and recognition systems that are in use today are rotation, pose, distance of recognition and illumination. These reduce the efficiency of the system unless performed under some necessary constraints. These constraints would involve positioning the subjects at specific positions, which in a real world classroom environment would be very hard and not to mention time consuming where the number of subjects involved is large.

With the help of a divergent combination of techniques and algorithms, this system helps us to achieve desired results with better accuracy. The provision of variable minimum detection scale eliminates the issue of distance for detection and recognition for both up close and group images. This has improved the face detection accuracy for upright frontal faces to 100% and consequently improved the face recognition accuracy from the typical efficiencies of 70%. Similarly, the minimum neighbors' setting has tremendously improved face detection accuracy.

Extracting and converting the rectangular part of the detected face instead of the whole image eliminates the effects of background noise on face detection improving the accuracy of the system. The camera in the system is used such that it only captures the frontal images so the problem of pose is not an issue. Histogram equalization is applied to the input images, this ensures that the output images are of uniform distribution of intensities through the reassignment of the intensity pixels. The

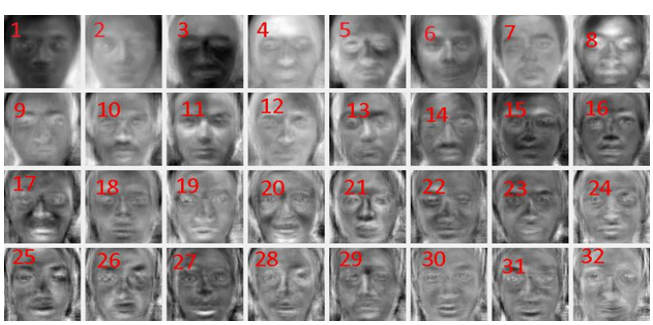


Fig. 14. *The first 32 Eigen faces.*

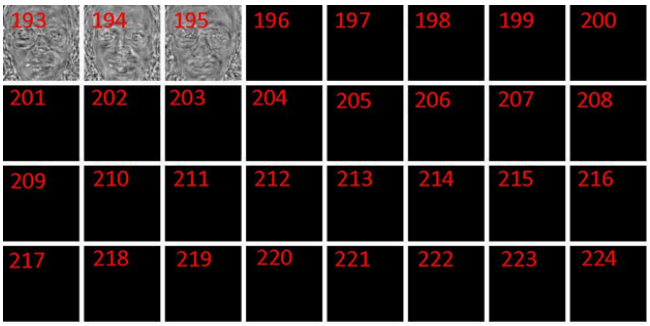


Fig. 15. *The last Eigen faces in the training set.*

input images of varying illumination are thus all enhanced in detail, this contributes into better face recognition results.

Fig. 14 shows the first 32 Eigen faces generated from a collection of 50 faces each of five people. The first few Eigen faces show dominant features of faces and the last Eigen faces from 196 to 247 are mainly image noise as shown in **Fig. 15** and are therefore discarded. The average face of **Fig. 16** obtained shows the smooth face structure of a generic human being.

From Figs. 14 and 15, it's seen that the first Eigen face shows the most dominant facial features of the training set images. The succeeding Eigen faces (principal components) in turn show the next highly probable facial features and more noise. Out of the 247 training images, 195 principal components together with the average face are enough to fully reconstruct the complete training set. We were therefore able to convert a set of correlated face variables (M) in to a set of values of K uncorrelated variables called principle components (eigenvectors). The number of Eigen faces is noted to be less than the original face images i.e. K < M.

Fig. 16. *The average face.*

From the Eigen faces obtained in the face recognition stage, it is interesting to discover that the principal components analysis can be used for image compression as evidenced by the dominant number of Eigen faces that can comfortably represent all images in the training set. Out of 247 images in the training set, only 195 faces together with the average faces are required to fully reconstruct the 247 faces in the set.

4. CONCLUSION

It can be concluded that a reliable, secure, fast and an efficient class attendance management system has been developed replacing a manual and unreliable system. This face detection and recognition system save time, reduce the amount of work done by the administration and replace the stationery material currently in use with already existent electronic equipment.

There is no need for specialized hardware for installing the system as it only uses a computer and a camera. The camera plays a crucial role in the working of the system hence the image quality and performance of the camera in real time scenario must be tested especially if the system is operated from a live camera feed.

The system can also be used in permission based systems and secure access authentication (restricted facilities) for access management, home video surveillance systems for personal security or law enforcement.

The major threat to the system is Spoofing. For future enhancements, anti- spoofing techniques like eye blink detection could be utilized to differentiate live from static images in the case where face detection is made from captured images from the classroom. From the overall efficiency of the system i.e. 83.1% human intervention could be called upon to make the system foolproof. A module could thus be included which lists all the unidentified faces and the lecturer is able to manually correct them.

REFERENCES

1. Shehu V, Dika A. Using real time computer vision algorithms in automatic attendance management systems. *32nd International Conference on Information Technology Interfaces*, Cavtat, Croatia, 2010.
2. Gopala M et al. Implementation of automated attendance system using face recognition. *International Journal of Scientific & Engineering Research*, 2015, 6(3).
3. Varadharajan E. Automatic attendance management system using face detection. *Online International Conference on Green Engineering and Technologies,* India, 2016.
4. Jadhav A, Jadhav A, Ladhe T, Yeolekar K. Automated attendance system using face recognition. *International Research Journal of Engineering and Technology*, 2017, 4(1):1467-1471.
5. Paharekari S, Jadhav C. Automated attendance system in college using face recognition and NFC. *International Journal of Computer Science and Mobile Computing*, 2017, 6(6):14-21.
6. Godswill O. Automated student attendance management system using face recognition. *International Journal of Educational Research and Information Science,* 2018, 5(4):31-37.

DOI: 10.17725/rensit.2020.12.305

Local heat source detection inside of the human body by means of microwave radiothermography

Evgeny P. Novichikhin
Kotelnikov Institute of Radioengineering and Electronics of RAS, Fryazino Branch, http://fire.relarn.ru/
Fryazino 141190, Moscow Region, Russian Federation
info@cplire.ru

Igor A. Sidorov
Radioengineering Concern "VEGA", http://www.vega.su/
Moscow 121170, Russian Federation
mail@vega.su

Vitaly Yu. Leushin
Scientific and Production Company «HIPERION», http://giperion-msk.ru/
Moscow 115201, Russian Federation
ra3bu@yandex.ru

Svetlana V. Agasieva
RUDN University, http://www.rudn.ru/
Moscow 117198, Russian Federation
s.agasieva@mail.ru

Sergey V. Chizhikov
Bauman Moscow State Technical University, https://bmstu.ru/
Moscow 105005, Russian Federation
chigikov95@mail.ru

Received August 02, 2019, reviewed August 22, 2019, accepted September 10, 2019

Abstract. **The possibility of non-invasive simultaneous detection of the depth and the temperature of a cancerous tumor inside the human body by means of multifrequency microwave 3D-radiothermograpy is regarding. The models for the description of the reception processes of the own human radio-thermal field are resulted. The possibility of calculating the required parameters by measuring antenna temperatures simultaneously in two different frequency ranges is analyzed. The conditions for solutions finding by bothanalytical and numerical methods are revealed. The possible maximum depth for tumors detection depending on the parameters of radiothermograph and thermal contrast in the source is determined. The necessitate of multi frequency receiving is approving. Analytical solutions for tumor depth and temperature for the current model are presented.**

Keywords: **microwave radiothermography, non-invasive temperature measurement, malignant tumor, 3D-visualization, antenna-applicator**

UDC 612.087

Acknowledgments. The study was carried out with a grant from the Russian Science Foundation (project No. 19-19-00349).

For citation: Evgeny P. Novichikhin, Igor A. Sidorov, Vitaly Yu. Leushin, Svetlana V. Agasieva, Sergey V. Chizhikov. Detection of a local source of heat in the depths of the human body by volumetric radiothermography. *RENSIT*, 2020, 12(2):305-312. DOI: 10.17725/rensit.2020.12.305.

Contents

1. INTODUCTION

Pathological processes inside the human body are usually accompanied by distortion of the natural heat field inside the body and on its surface. Knowledge of the heat field distribution in the human body and the reaction of the heat field to various physiological tests allows reliably to diagnose various diseases. The external temperature of the human body is measured by conventional medical thermometers or infrared pyrometers and thermal imagers. It is impossible to measure the temperature inside the body by such methods, and the introduction of a thermal sensor under the skin leads to a violation of the natural heat field.

Therefore, it is urgent to improve non-invasive methods of measuring internal temperatures in the human body for the purpose of early diagnosis and monitoring of malignant neoplasms and other pathologies by radiothermography, which is actively developed by specialists and scientists all over the world [1].

Some scientists have hypothesized that a long inflammatory process can eventually lead to a malignant neoplasm. Traditional diagnostic methods (magnetic resonance imaging (MRI), computed tomography (CT), etc.) give the doctor information about structural changes in tissues: the size of the tumor, its localization, the presence of microcalcinates, density, and allow to identify, mainly, already formed tumors at "clinically late" stages of development. Temperature is the first marker of pathological changes in the human body. For example, the temperature of a malignant tumor due to increased metabolism is 2-3 degrees higher than the temperature of intact tissues. Moreover, thermal changes occur not only when there is a high probability of malignancy. You can get information about the temperature of internal tissues using MRI, but this approach requires access to sophisticated medical equipment. MRI equipment has a high cost and is not suitable for measurements that need to be repeated frequently over a long period of time. It opens up huge opportunities for applying the radiometry method in practical medicine [7-9].

However, the development of this method is hindered by the presence of a number of scientific and technical barriers that need to be overcome. Combining in one radiometric complex the principles of multichannel, multi-frequency and microminiature will lead to a significant reduction in the size of the radiometric receiver and the need to develop fundamentally new design and technological solutions, namely, its implementation in the form of a single module, which implies the use of a monolithic integrated design. The results of work in this direction are shown in the works [10-16].

Another main problem that the research has described in this article is that the construction of 3*D* images of radio-brightness temperatures based on electromagnetic radiation registered by a digital module for processing radiometric signals built on new principles requires the development of a fundamentally new set of algorithms and programs that are adequated to the biological object under study.

The radiothermography method is based on receiving and measuring the characteristics of the human body's own radiothermal radiation using a specialized high – sensitivity receiver in the range of centimeter or decimeter waves-a microwave radiometer with special antenna applicators [2] installed on the surface of the human body. At the same, for ensuring an acceptable accuracy of temperature measurements (the order of 0.1 degrees), it is necessary to take into account the degree of coordination of the antenna applicators with the human body at the installation places, which is achieved due to a special reception mode-scatterometric reception.

It is especially effective to use radiothermography for malignancies (cancers)

detection in the early stages of the development of pathology, even when they are not yet detected by *X*-ray method. The radiothermography method is also applicable for glucose testing, when the patient is given to drink 30 grams of an aqueous glucose solution on an empty stomach. Glucose, as a high-calorie substance, is absorbed and carried by the bloodstream throughout the body, feeding cells [3]. At the same time, the body temperature increases uniformly for a short time by one to two tenths of a degree. If there is a malignancy somewhere, then the temperature increases significantly more at the place of its localization, by one or two degrees. Detecting of a local temperature anomaly shows where the cancer is located.

Multichannel radiothermography enable to receive, process and visualize data from multiple antennas-applicators at the same time. This displays a two-dimensional image of the temperature distribution, which changes over time during the analysis [4]. However, this method does not allow to determine the depth of the tumor location under the skin.

The purpose of this article is to show the possibility of determining not only the location of the cancer, but also the depth of cancer location using volumetric radiothermography.

2. METHODOLOGICAL BASIS OF 3D RADIOTHERMOGRAPHY

The method of volumetric radiothermography is based on the use of natural electromagnetic radiation from various objects (including tissues of living creatures), whose temperature is different from absolute zero [16-17]. Any element of the human body is a source of thermal electromagnetic radiation in a wide range of frequencies. Radiation that occurs in the depth of the human body, spreading to the surface, is partially damped by absorption in human tissues. The amount of wave attenuation depends on the type of tissue (muscle, fat, bone, cranial, brain) and the wavelength. Numerically, attenuation is characterized by the size of the skin layer or the depth at which the power of the electromagnetic wave decreases by a factor of e (2.7282). The size of the skin layer depends on the wavelength of radiation. So, for a 43 cm wave, the value of the skin layer for breast tissue is about 7 cm, and for a 21 cm wave-about 3.5 cm. Thus, by measuring the power of the body's own thermal radiation at one point, but in different frequency ranges, it is possible to differentiate the location of the source of increased thermal radiation by depth. This is the methodological basis of three dimensional radiothermography.

For refining the algorithms of processing signals received by a multi-channel multi-frequency radiothermograph and calculating the temperature distribution within the human body by depth, it is necessary mathematical modeling of heat transfer processes and thermal fields to calculate the values of radio-brightness temperatures received by the radiothermograph in each frequency range [5]. In practice, it is necessary to solve the inverse problem, namely, to calculate the distribution of thermodynamic temperature by depth for each applicator antenna from the measured values of radio-brightness temperatures at different frequency ranges and from the measured values of surface temperatures. The obtained values must be interpreted over the entire surface of the body and over the depth to restore the 3*D* structure of the thermal field. Analysis of the 3*D* dynamic picture of the structure of the heat field allows us to determine three coordinates of the local source of abnormal heating, if it is present, which will allow us to more accurately localize the position of the malignant tumor.

Using of several antennas and multi-channel radiometric receivers operating in the microwave range allows to make dynamic studies of deep human body temperatures with computer processing and presentation of results in the form of temperature maps and dynamic graphs. Providing the required resolution and sensitivity in real time is an extremely difficult task. It will be possible to use an affordable and inexpensive

device for early diagnosis of a large number of pathologies for personal medicine. It should be particularly noted that in addition to early diagnosis of various pathologies, it can be used for non-invasive monitoring of the disease treatment process [19-20].

A conventional radiothermograph [6] measures the average temperature from the radiation pattern of the applicator antenna inside the human body. The novelty of this approach consists in an attempt to more accurately locate a point heat source in the main beam area of the antenna-applicator directional diagram and calculate its temperature.

3. MODELING OF HEAT FIELDS AND HEAT TRANSFER PROCESSES IN THE HUMAN BODY

A complete model of the human body that takes into account all processes (heat release, heat transfer, radiation of the thermal electromagnetic field, its propagation and reception) is extremely cumbersome and complex. We consider several models built on the principle of "from simple to complex" to solve this problem.

A section of the human body is considered as a homogeneous medium of electromagnetic waves propagation with a constant absorption coefficient and without thermal conductivity. The applicator antenna, that is perfectly aligned with the body at the installation place, has a pencil form of directivity inside the human body, and does not have side lobes or back scattering. The cancer is a point source of heat with an increased temperature compared to the body temperature. We can consider the area with the tumor an absolutely black body, then its radio-brightness temperature is equal to the thermodynamic one without limiting generality. In practice, the tumor is a "grey" body, since it has not yet been detected differences in the dielectric properties of normal tissues and tissues affected by the tumor. Considering the tumor as a "grey" body does not limit the generality of the model, but only leads to a decrease in brightness contrast. The location of the model elements is shown in **Fig. 1**. The model is considered in a coordinate system with the beginning at the point of installation of the antenna-applicator, the *X* and *Y* axes in the plane of the skin surface, the *Z* axis is directed from the surface to the depth of the human body. All further calculations will be considered in one-dimensional space along the *Z*-axis.

General, the brightness temperature is defined as follows (1):

$$T_b = \int_0^\infty w(z)T(z)dz, \quad (1)$$

where $T(x)$ - thermodynamic temperature, $w(x)$ - the weight function determined by absorption, and

$$\int_0^\infty w(z)dz = 1.$$

As the absorption changes by the exponential law:

$$w(z) = ke^{-kz},$$

where k - absorption coefficient for a given wavelength. Value, reverse k is a value for the skin layer z_s, the thickness of the layer at which the radiation decreases in e times.

$$z_s = \frac{1}{k}.$$

Let the tumor have a temperature T_c located at a depth of z_s, and T_0 - body temperature. Given

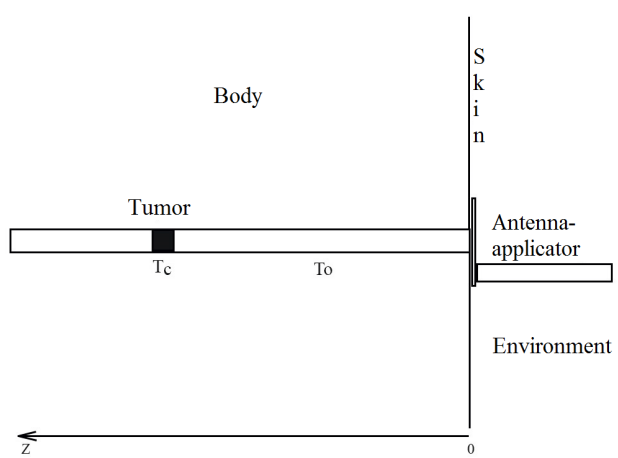


Fig. 1. *Model for locatization of a heat source in the human body.*

that the tumor is assumed to be a completely black body, the temperature distribution in the body can be written as a function:

$$T(z)=\begin{cases}T_0, \ if \ z<z_c,\\ T_c, \ if \ z\geq z_c.\end{cases} \quad (2)$$

Then, substituting this expression in equation (1), we get the brightness temperature measured by a microwave radiometer connected to the antenna-applicator:

$$T_b = T_0\rho + T_c(1-\rho), \quad (3)$$

where

$$\rho = 1-e^{-kz_c} = 1-e^{-\frac{z_c}{z_s}}. \quad (4)$$

Substituting formula (4) in formula (3) we get:

$$T_b(z_c) = (T_c - T_0)e^{-\frac{z_c}{z_s}} + T_0. \quad (5)$$

Formula (5) is valid for all frequency channels in the decimeter range, but the value of the z_s skin layer depends on the frequency range. The graph of the antenna temperature dependence calculated by the formula (5) is shown in **Fig. 2** by solid line.

Formula (5) allows us to estimate the maximum depth at which a source with temperature T_c can be detected by a radiometer with sensitivity δT:

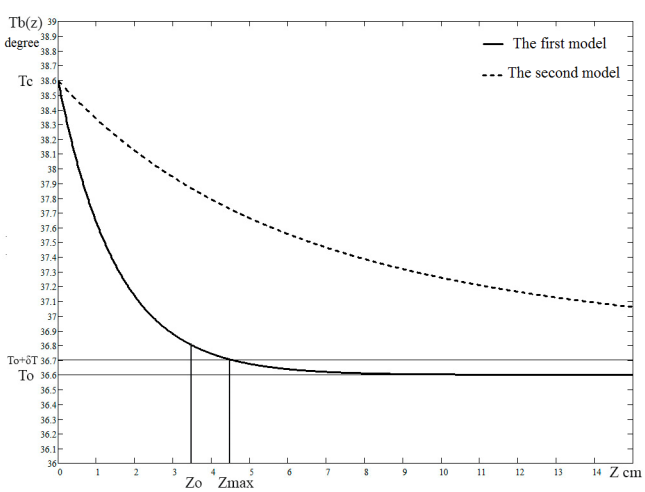


Fig. 2. *Brightness temperature values for two models calculated using formulas (3) and (11). The depth of the tumor is deposited in cm on the x-axis for the skin layer - 3.5 cm.*

$$z_{\max} = z_s \cdot \ln\left(\frac{T_C - T_0}{\delta T}\right). \quad (6)$$

Analysis of formula (6) shows that the maximum detection depth of a heat source directly depends on the value of the skin layer (attenuation in the medium, than attenuation is less that the detection depth is greater), depends on the value of the thermal contrast (the difference between the source temperature and body temperature) and the sensitivity of the radiometer (than the sensitivity is higher, that the detection depth is greater).

In formula (5), the known values are $T_b(z_s)$ (measured by a radiometer), the body temperature T_0 (measured by a thermal sensor on the surface of the body), the value of the skin layer for a given frequency range zs, can be determined by special calibration. Unknown are the temperature of the source T_c and the depth of its occurrence zc. One equation with two unknowns does not have an unambiguous solution, therefore, a single-frequency (single-band) radiothermograph is not able to simultaneously measure the temperature and depth of the tumor.

Measurements must be made simultaneously in at least two different frequency ranges with frequencies λ_1 and λ_2 to uniquely determine the temperature of the tumor and the depth of its occurrence. Then applying the formula (5) to each frequency range, we get a system of two equations with two unknowns:

$$\begin{cases} T_{b\lambda_1}(z_c) = (T_c - T_0)e^{\frac{z_c}{z_{s\lambda_1}}} + T_0 \\ T_{b\lambda_2}(z_c) = (T_c - T_0)e^{\frac{z_c}{z_{s\lambda_2}}} + T_0 \end{cases}. \quad (7)$$

It should be noted that the solution of the system of equations (7) makes sense to determine only under the condition that the source T_c will find both frequency channels, more accurately, you find the shortwave channel, as if the source is found a shortwave channel, it will be found by a longwave channel. If this condition is, the system of equations (7) has an unambiguous analytical solution:

$$z_c = \frac{z_{s\lambda_1} z_{s\lambda_2}}{z_{s\lambda_1} - z_{s\lambda_2}} \ln\left(\frac{T_{b\lambda_1} - T_0}{T_{b\lambda_2} - T_0}\right)$$
$$T_c = T_0 + \frac{T_{b\lambda_2} - T_{b\lambda_1}}{\alpha^{z_{s\lambda_1}} - \alpha^{z_{s\lambda_2}}}, \quad (8)$$

where

$$\alpha = \left(\frac{T_{b\lambda_1} - T_0}{T_{b\lambda_2} - T_0}\right)^{\frac{1}{z_{s\lambda_1} - z_{s\lambda_2}}}.$$

Analysis of formulas (8) shows that expressions make sense if $z_{s\lambda_1} \neq z_{s\lambda_2}$,, which always holds under the condition that $\lambda_1 \neq \lambda_2$. And the conditions $T_{b\lambda_1} \neq T_0$ and $T_{b\lambda_2} \neq T_0$, must also be, which means that the T_c source can be detected in both frequency channels, as noted above.

A more accurate model should take into account the distribution of heat and, consequently, the temperature inside the body. The classical method consists in solving the heat equation, but given the complexity and heterogeneity of the object, the presence of heat transfer by blood, etc. this can only be done by numerical methods.

Therefore, a one-dimensional stationary equation of thermal conductivity in a homogeneous medium is considered as the next step. In this case, the solution is represented by a linear equation in the area from *0* to z_c (from the surface to the tumor). Taking into account the previous assumptions, the temperature function can be represented as:

$$T(z) = \begin{cases} az + b, \ z < z_c, \\ T_c, \ z \geq z_c. \end{cases} \quad (9)$$

The values of parameters *a* and *b* are determined by the boundary conditions for the section *(0,z)*, namely

$$\begin{cases} T_0 = a0 + b = b \\ T_c = az_c + b \end{cases} \Rightarrow \begin{cases} a = \dfrac{T_c - T_0}{z_c} \\ b = T_0 \end{cases}. \quad (10)$$

In general case, the boundary condition for $z = 0$ is set by conditions of the 3rd kind, i.e. by the heat flow between the surface and the external environment, but taking into account the fact that the temperature on the surface is set, i.e. always measurable, a simpler option is chosen.

Substituting the temperature function in the expression for the brightness temperature and taking into account the boundary conditions, we get:

$$T_b = T_0\rho' + T_c(1 - \rho'), \quad (11)$$

where

$$\rho' = 1 - \frac{z_s}{z_c}(1 - e^{-\frac{z_c}{z_s}}).$$

It can be seen that equation (11) is identical to equation (3) for a simpler model. The difference is that the system of equations for two frequencies does not have an analytical solution in this case.

Calculations of the brightness temperature T_b for two models are given in Fig. 2. So the body temperature T_0 is 36.6 degrees, and the tumor temperature T_c -38.6 degrees. It can be seen that in the case of model 2, the effect of the tumor on the measured value of the brightness temperature is expected to be greater than in model 1, since it takes into account the distribution of heat in the body.

4. CONCLUSION

New results were obtained as a result of the research and modeling within the framework of the built models:

- it was shown that using a single-frequency radiothermograph it is impossible to simultaneously determine the temperature of a local thermal anomaly and the depth of its occurrence. There was a justification for the need to make measurements in at least two different frequency ranges and use data from surface thermal sensors to determine the temperature of a local thermal anomaly and the depth of its occurrence;
- there were proposed 2 models of brightness temperature formation on the body surface;
- the maximum depth of thermal anomaly detection in the human body was determined

depending on the value of the skin layer for a given wavelength, thermal contrast, and the sensitivity of the radiometer;
- it was obtained a system of equations describing the dependences of the measured and desired physical quantities;
- it was shown that the solution of the system of equations is possible only if the thermal anomaly occurs at a depth not exceeding the maximum detection depth of the thermal anomaly for a shorter frequency channel of the radiothermograph;
- analytical solutions for the temperature of a thermal anomaly and its depth were obtained for one of the models.

REFERENCES

1. Gulaev YV, Verba VS, Gandurin VA, Gudkov AG, Leushin VY, Tsiganov DI. Passivnye i aktivnye radiolokatsionnye metody issledovanij i diagnostika zhivykh tkanej cheloveka [Passive and active radar methods of research and diagnostics of human living tissues]. *Biomedicinskie tehnologii i radio elektronika,* 2006, 11:14-20.
2. Vesnin SG, Sedankin MK. Miniatyurnye antenny appplikatory dlya microvolnovykh radiometrov meditsinskogo naznacheniya [Miniature antenna-applicator for medical microwave radiometers]. *Biomedicinskaya radioelektronika,* 2011, 10:51-56.
3. Gautherie M. Temperature and blood flow patterns in breast cancer during natural evolution and following radiotherapy. *Progress in Clinical and Biological Reserch,* 1982, 107:21-64.
4. Gulaev UV, Leushin VY, Gudkov AG, Schukin SI, Vesnin SG, Kublanov VS, Porohov IO, Sedankin MK, Sidorov IA. Pribory dlya diagnostiki patologicheskikh izmenenij v organizme cheloveka metodami mikrovolnovoj radiometrii [Devices for pathological changes diagnosis in the human body by means of microwave radiometry]. *Nanjtehnologii: razrabotka, primenenie – XXI vek,* 2017, 9(2):27-45.
5. Sedankin MK, Leushin VYu, Gudkov AG, Vesnin SG, Sidorov IA, Agasieva SV, Markin AV. Mathematical Simulation of Heat Transfer Processes in a Breast with a Malignant Tumor. *Biomedical Engineering,* 2018, 52(3):190-194. DOI: 10.1007/s10527-018-9811-2.
6. Birukov ED, Verba VS, Gudkov AG, Leushin VY, Plushchev VA, Sidorov IA. Multi-frequency radiothermograph. *Patent RF RU2328751*. Priority 20.02.2008. IPC Class G01R29/08.
7. Verba VS, Gandurin VA, Gudkov AG, Leushin VY, Plushev VA. Application of radiolocation methods in radio frequency and optical bands to detect human living tissue pathologies. *Proc. 16th Intern. Crimean Conf Microwave and Telecommunication Technology (CriMiCo-2006)*, № 4023532:903-904.
8. Gudkov AG, Leushin VY, Mikheev VA, Kolpakov NS, Porokhov IO, Silkin AT. The set of wideband directional antennas for systems of electromagnetic radiation locating. *Proc. 21st Intern. Crimean Conf. Microwave and Telecommunication Technology* (*CriMiCo-2011*), № 6069055:561-562.
9. Agasieva SV, Gudkov AG, Korolev AV, Leushin VY, Plushchev VA, Sidorov IA Development results of the unified receiving module for multichannel medical radio thermographs.*Proc. 24th Intern. Crimean Conf. Microwave and Telecommunication Technology (CriMiCo-2014*), № 6959752:1045-1046.
10. Gudkov AG. Complex technological optimization of microwave devices. *Proc. 17th Intern. Crimean Conf. Microwave and Telecommunication Technology* (*CriMiCo-2007)*, № 4368833:521-522.
11. Gudkov AG, Leushin VY, Meshkov SA. Ensuring of microwave circuit quality performance using the method of probablistic modeling. *Proc. 17th Intern. Crimean Conf. Microwave and Telecommunication*

Technology (*CriMiCo-2007)*, № 4368834:523-524.

12. Gudkov AG, Leushin VY, Meshkov SA, Popov VV. Application of complex technological optimization for monolithic microwave circuits designing. *Proc. 18th Intern. Crimean Conf.Microwave and Telecommunication Technology (CriMiCo-2008*), № 4676491:535-536.
13. Tikhomirov VG, Gudkov AG, Agasieva SV, Gorlacheva EN, Shashurin VD, Zybin AA, Evseenkov AS, Parnes YM. The sensitivity research of multiparameter biosensors based on HEMT by the mathematic modeling method. *Journal of Physics: Conference Series*, 2017, 917(4):042016.
14. Parnes YM, Tikhomirov VG, Petrov VA, Gudkov AG, Marzhanovskiy IN, Kukhareva ES, Vyuginov VN, Volkov VV, Zybin AA. Evaluation of the influence mode on the CVC GaN HEMT using numerical modeling. *Journal of Physics: Conference Series,* 2016, 741 (1):012024.
15. Sister VG, Ivannikova EM, Gudkov AG, Leushin VY, Sidorov IA, Plyushchev VA, Soldatenko AP. Detection of forest and peat-bog fire centers by means of microwave radiometer sounding. *Chemical and petroleum engineering*, 2016, 52(1-2):123-125.
16. Gudkov AG, Sister VG, Ivannikova EM, Sidorov IA, Chetyrkin DY. On the Possibility of Detecting Oil Films on a Water Surface by Methods of Microwave Radiometry. *Chemical and Petroleum Engineering*, 2019, 55(1-2):57-62.
17. Sedankin MK, Leushin VY, Gudkov AG, Agasieva SV, Markin AV etc. Mathematical Simulation of Heat Transfer Processes in a Breast with a Malignant Tumor. *Biomedical Engineering,* 2018, 52(3):190-194.
18. Sedankin MK, Leushin VY, Gudkov AG, Agasieva SV, Gorlacheva EN. Modeling of Thermal Radiation by the Kidney in the Microwave Range. *Biomedical Engineering,* 2019, 53(1):60-65.
19. Sedankin MK, Leushin VY, Gudkov AG, Ovchinnikov LM, Vetrova NA. etc. Antenna Applicators for Medical Microwave Radiometers. *Biomedical Engineering,* 2018, 52(4):235-238.
20. Bobrikhin AF, Gudkov AG, Leushin VY, Los' VF, Porokhov IO, Sidorov IA. Modeling of the dipole, helical and cavity-slot antennas applicators for multichannel medical radiothermographs. *Proc. 24th Intern. Crimean Conf. Microwave and Telecommunication Technology* (*CriMiCo 2014*), № 6959753:1047-1048.

Radioelectronics. Nanosystems. Information Technologies (abbr. RENSIT)
Certificate El. no. FS77-60275 on 19.12.2014 of the Ministry of Telecom and Mass Communications of Russian Federation, Moscow
ISSN: 2414-1267online. Http://www.rensit.ru. Publisher - Vladimir I. Grachev. 2020, 30 August.
Computer printing, page-proofs, graphics, photos of work - the editors RENSIT.