APPLIED NUMERICAL MATHEMATICS and SCIENTIFIC COMPUTATION

Proceedings of the 2nd International Conference on Applied Mathematics and Computational Methods (AMCM 2014)

Proceedings of the 2nd International Conference on Mechanics, Fluids, Heat, Elasticity and Electromagnetic Fields (MFHEEF 2014)

Proceedings of the 1st International Conference on Civil Engineering, Water Resources, Hydraulics & Hydrology (CEWHH 2014)

Proceedings of the 2nd International Conference on Biology, Medical Physics, Medical Chemistry, Biochemistry and Biomedical Engineering (BIOMED 2014)

Proceedings of the 1st International Conference on Chemistry, Chemical Engineering and Materials Science (CCEMS 2014)

Proceedings of the 1st International Conference on Theoretical and Applied Physics (TAP 2014)

Athens, Greece
November 28-30, 2014

APPLIED NUMERICAL MATHEMATICS and SCIENTIFIC COMPUTATION

Proceedings of the 2nd International Conference on Applied Mathematics and Computational Methods (AMCM 2014)
Proceedings of the 2nd International Conference on Mechanics, Fluids, Heat, Elasticity and Electromagnetic Fields (MFHEEF 2014)
Proceedings of the 1st International Conference on Civil Engineering, Water Resources, Hydraulics & Hydrology (CEWHH 2014)
Proceedings of the 2nd International Conference on Biology, Medical Physics, Medical Chemistry, Biochemistry and Biomedical Engineering (BIOMED 2014)
Proceedings of the 1st International Conference on Chemistry, Chemical Engineering and Materials Science (CCEMS 2014)
Proceedings of the 1st International Conference on Theoretical and Applied Physics (TAP 2014)

Athens, Greece
November 28-30, 2014

All papers of the present volume were peer reviewed by no less than two independent reviewers. Acceptance was granted when both reviewers' recommendations were positive.

ISBN: 978-1-61804-253-8

APPLIED NUMERICAL MATHEMATICS and SCIENTIFIC COMPUTATION

Proceedings of the 2nd International Conference on Applied Mathematics and Computational Methods (AMCM 2014)

Proceedings of the 2nd International Conference on Mechanics, Fluids, Heat, Elasticity and Electromagnetic Fields (MFHEEF 2014)

Proceedings of the 1st International Conference on Civil Engineering, Water Resources, Hydraulics & Hydrology (CEWHH 2014)

Proceedings of the 2nd International Conference on Biology, Medical Physics, Medical Chemistry, Biochemistry and Biomedical Engineering (BIOMED 2014)

Proceedings of the 1st International Conference on Chemistry, Chemical Engineering and Materials Science (CCEMS 2014)

Proceedings of the 1st International Conference on Theoretical and Applied Physics (TAP 2014)

Athens, Greece
November 28-30, 2014

Organizing Committee

Additional Reviewers

Eleazar Jimenez Serrano	Kyushu University, Japan
Xiang Bai	Huazhong University of Science and Technology, China
Jose Flores	The University of South Dakota, SD, USA
Genqi Xu	Tianjin University, China
Konstantin Volkov	Kingston University London, UK
João Bastos	Instituto Superior de Engenharia do Porto, Portugal
Abelha Antonio	Universidade do Minho, Portugal
Miguel Carriegos	Universidad de Leon, Spain
Tetsuya Yoshida	Hokkaido University, Japan
Bazil Taha Ahmed	Universidad Autonoma de Madrid, Spain
Moran Wang	Tsinghua University, China
Yamagishi Hiromitsu	Ehime University, Japan
Philippe Dondon	Institut polytechnique de Bordeaux, France
Manoj K. Jha	Morgan State University in Baltimore, USA
Frederic Kuznik	National Institute of Applied Sciences, Lyon, France
Minhui Yan	Shanghai Maritime University, China
Lesley Farmer	California State University Long Beach, CA, USA
Zhong-Jie Han	Tianjin University, China
Stavros Ponis	National Technical University of Athens, Greece
Ole Christian Boe	Norwegian Military Academy, Norway
Imre Rudas	Obuda University, Budapest, Hungary
Hessam Ghasemnejad	Kingston University London, UK
Matthias Buyle	Artesis Hogeschool Antwerpen, Belgium
Kazuhiko Natori	Toho University, Japan
Dmitrijs Serdjuks	Riga Technical University, Latvia
George Barreto	Pontificia Universidad Javeriana, Colombia
Kei Eguchi	Fukuoka Institute of Technology, Japan
James Vance	The University of Virginia's College at Wise, VA, USA
Shinji Osada	Gifu University School of Medicine, Japan
Francesco Rotondo	Polytechnic of Bari University, Italy
Valeri Mladenov	Technical University of Sofia, Bulgaria
M. Javed Khan	Tuskegee University, AL, USA
Andrey Dmitriev	Russian Academy of Sciences, Russia
Angel F. Tenorio	Universidad Pablo de Olavide, Spain
Jon Burley	Michigan State University, MI, USA
Deolinda Rasteiro	Coimbra Institute of Engineering, Portugal
Sorinel Oprisan	College of Charleston, CA, USA
Francesco Zirilli	Sapienza Universita di Roma, Italy
Alejandro Fuentes-Penna	Universidad Autónoma del Estado de Hidalgo, Mexico
Tetsuya Shimamura	Saitama University, Japan
Masaji Tanaka	Okayama University of Science, Japan
Takuya Yamano	Kanagawa University, Japan
Santoso Wibowo	CQ University, Australia
José Carlos Metrôlho	Instituto Politecnico de Castelo Branco, Portugal

Table of Contents

ISBN: 978-1-61804-253-8

ISBN: 978-1-61804-253-8

ISBN: 978-1-61804-253-8

ISBN: 978-1-61804-253-8

Necessary Optimality Conditions for Parabolic Equations with Venttsel Boundary Control

Yousong Luo
School of Mathematical and Geospatial Sciences,
RMIT University, GPO Box 2476V Melbourne, Vic. 3001, AUSTRALIA
email: yousong.luo@rmit.edu.au

***Abstract*—In this paper we study the optimality condition for the Venttsel boundary control of a parabolic equation, that is, the state of the dynamic system is governed by a parabolic equation together with an initial condition while the control is applied to the system via the Venttsel boundary condition. The first and second order necessary conditions are derived for the optimal solution in the case of both unconstrained and constrained problems.**

***Index Terms*—Optimality condition, Parabolic equation Vettsel boundary condition, Boundary control**

I. Introduction

In this paper we discuss the necessary optimality conditions for a class of optimal control problems formulated as follows:

$$\begin{cases} \text{Minimize } \ J(u) \\ u \in U \end{cases} \tag{1}$$

where u is the control chosen from an allowable set U, y_u is the output state variable governed by a state equation corresponding to the input u, $J(u)$ is the objective function.

The state equation under our consideration is an initial-boundary value problem of parabolic equations where the control u is applied to the dynamic system via the Venttsel boundary condition. We will give the detailed introduction of the state equation, objective and constrain functions in Section 2.

The optimal control problems of systems governed by partial differential equations are well studied. The second order necessary and sufficient optimality conditions for elliptic problems are obtained by Casas E. and Tröltzsch F. in [7] and [8]. Similar conditions for parabolic problems are studied by Raymond J. and Tröltzsch F. in [17], Krumbiegel K. and Rehberg J. in [11]. The general theory on PDE control problems can be found in standard textbooks such as [12] or Raymond on-line lecture notes. In those literatures, either the interior distributed control or boundary controls through Dirichlet, Neumann and general oblique boundary conditions are well studied. This paper will contribute the Venttel boundary control to the existing theory in this field.

An initial-boundary value problem of a parabolic equation with a parabolic Venttsel boundary condition arises in the engineering problem of heat conduction. A simple example is the problem of heat conduction in a medium enclosed by a thin skin and the conductivities of the medium and the surrounding skin are significantly different, see [6] and [16]. Generally speaking, all physical phenomena involving a diffusion process along the boundary manifold will give rise to a Vettsel type boundary condition as it gives rise to a second order tangential derivatives (diffusion) as well as the first and zero order derivatives the unknown function. The theoretical frame work in dealing with such a boundary problems has been developed since 1990's. It was started with elliptic equations by Luo Y. and Trudinger N. in [14] and [15] and continued with parabolic equations by Apushkinskaya D. and Nazarov A. in [1] , [2] and [3]. The existence, uniqueness as well as the *a priori* estimates of both classical and distributional solutions are established. It has been shown in [18] that the Venttsel boundary condition is the most general feasible boundary condition for a parabolic or elliptic equation and, in the degenerate case where the second order term vanishes, it includes Dirichlet, Neumann and general oblique boundary conditions as special cases.

For the optimal control problems involving Venttsel boundary condition, a first order necessary condition is derived by the author in [13] where the state equation is an elliptic equation, based on the results in [4], [5] and [7]. This paper is the continuation of [13] and it is also an analogue to [7], [8] and [17] because similar results have already been obtained for elliptic problems or parabolic problems with traditional boundary conditions.

In the following, we will first state the problems clearly and collect all relevant back ground results for the solutions of our state equation in Section 2. In Section 3 we will establish the differentiability of the objective functional and derive a formula to express the derivatives of it. In Section 4, we will give the optimal condition for both unconstrained and constrained problems. Finally in Section 5 we will make some comments on further development.

II. Preliminary

A. Notation, function spaces

Let Ω be a bounded open subset of $\mathbb{R}^n$ with a C^3 boundary $\Gamma = \partial\Omega$. For $T > 0$, we define $Q = \Omega \times (0,T)$ and $\Sigma = \Gamma \times (0,T)$. For functions $y : Q \to \mathbb{R}$ the notation $D_i y$ denotes the partial derivative with respect to the space variable x_i and $D_t y$ denotes the partial derivative with respect to the time variable t. $Dy = (D_1 y, \ldots, D_n y)$ is the gradient of y. We also denote by

$$D^\sigma y := \frac{\partial^{|\sigma|} y}{\partial x_1^{\sigma_1} \partial x_2^{\sigma_2} \cdots \partial x_n^{\sigma_n}}$$

ISBN: 978-1-61804-253-8

the partial derivatives of order $|\sigma|$ where $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_n) \in \mathbb{N}_0^n$ is a multi-index of modulus $|\sigma| = \sum_{i=1}^n \sigma_i$. The parabolic distance between the points $P_1(x_1, t_1)$ and $P_2(x_2, t_2)$ in Q is defined by $d(P_1, P_2) = (|x_1 - x_2|^2 + |t_1 - t_2|)^{1/2}$ where $|x|$ is the Euclidean norm in $\mathbb{R}^n$. When a point P is on the boundary Σ we usually write its coordinate as (s, t) where s is a variable on Γ and if an integral is involved we will use ds to denote the surface area element of Γ.

The space $C(\overline{Q})$ is the Banach space of all continuous functions y in $\overline{Q}$ with the norm

$$\|y\|_Q = \sup_Q |y|.$$

For $0 < \alpha < 1$, the space $C^\alpha(\overline{Q})$ is the Banach space of functions with the norm

$$\|y\|_{C^\alpha(\overline{Q})} = \|y\|_Q + [y]_{\alpha,Q}$$

where $[\cdot]_{\alpha,Q}$ stands for the Hölder semi-norm

$$[y]_{\alpha,Q} := \sup_{P_1,P_2 \in Q} \frac{|y(P_1) - y(P_2)|}{d(P_1, P_2)^\alpha}.$$

The space $C^{2,\alpha}(\overline{Q})$ is the Banach space of functions with the norm

$$\|y\|_{C^{2,\alpha}(\overline{Q})} = \sum_{|\sigma| \le 2} \|D^\sigma y\|_Q + \|D_t y\|_Q + \sum_{|\sigma|=2} [D^\sigma y]_{\alpha,Q} + [D_t y]_{\alpha,Q}.$$

The restriction of $C^{2,\alpha}(\overline{Q})$ functions on the boundary of Σ is denoted by $C^{2,\alpha}(\Sigma)$. When the functions are independent of t we can define the spaces $C^{2,\alpha}(\overline{\Omega})$ and $C^{2,\alpha}(\Gamma)$ in exactly the same way. Notice also that every $C^{2,\alpha}(\Sigma)$ function can always be extended to a $C^{m,\alpha}(\overline{Q})$ function and such an extension can be carried out in a manner that preserves the norm, i.e. the corresponding $C^{2,\alpha}(\Sigma)$ norm and $C^{2,\alpha}(\overline{Q})$ are equivalent. Based on such an observation we will not distinguish the spaces $C^{2,\alpha}(\Sigma)$ and $C^{2,\alpha}(\overline{Q})$.

Let $\nu = (\nu^1, \ldots, \nu^n)$ be the outward unit normal vector field of Γ. Then the outward normal derivative of y, denoted by $\partial_\nu y$, is defined by

$$\partial_\nu y = Dy \cdot \nu$$

where Dy is the gradient vector of y. Now we define the tangential differential operators. Let $\{c^{ik}\}_{n\times n}$ be the matrix whose entries are given by

$$c^{ik} = \delta^{ik} - \nu^i \nu^k,$$

where δ^{ik} is the Kronecker symbol. Then the first and the second order tangential differential operators are then defined by

$$\partial_i = c^{ik} D_k, \quad \partial_{ij} = \partial_i \partial_j, \quad i, j, = 1, \cdots, n,$$

hence the tangential gradient operator is defined by

$$\partial = (\partial_1, \cdots, \partial_n).$$

In particular the Laplace-Beltrami operator on the boundary manifold is then defined by

$$\Delta_\Gamma = \partial_i \partial_i.$$

All repeated indecis above indicate a summation from 1 to n. Note that the second order tangential derivatives so defined are not symmetric in general.

B. State equations, objective functionals and constraints

The state equation in this paper is the following semi-linear initial-boundary value problem of heat equation

$$\text{(SE)} \quad \begin{cases} D_t y - \Delta y = f & \text{in } Q, \\ D_t y - \Delta_\Gamma y + \partial_\nu y = \sigma y + \varphi(s,t,u), & \text{on } \Sigma, \\ y(x,0) = y_0 & \text{in } \Omega. \end{cases} \tag{2}$$

where $f \in C^\alpha(Q)$ and $\sigma \in C^\alpha(\Sigma)$ are given functions, $y_0 \in C^{2,\alpha}(\Omega)$ is the initial temperature, $u \in C^\alpha(\Sigma)$ is the control function and φ is a given smooth function. We will not precisely specify the class that φ belongs to and assume that all derivatives needed exist and are bounded as long as all the variables in φ are bounded.

The boundary condition of this kind is known as the Venttsel boundary condition. As mentioned in the introduction, physically the Venttsel boundary condition occurs when the boundary manifold Γ and the domain Ω have significantly different conductivity. In such a case the boundary condition should take the form of

$$D_t y - \kappa \Delta_\Gamma y + \partial_\nu y = \sigma y + \varphi(s,t,u)$$

where κ is a positive constant not equal to 1 if we normalize the heat equation in Q to the form in (2). However this does not cause any difference in the following theoretical development. Due to such a reason we only consider the state equation in the form of (2).

The objective functional $J : C^\alpha(\Sigma) \to \mathbb{R}$ is given by

$$J(u) = \int_Q p(x,t,y_u)\, dxdt + \int_\Sigma q(s,t,y_u,u)\, dsdt \tag{3}$$

where $p : Q \times \mathbb{R} \to \mathbb{R}$ and $q : \Sigma \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are of class C^1 and $y_u = G(u)$ is the solution of the state equation (2) corresponding to the control u.

Let u_a and u_b be a given pair of functions in $C^\alpha(\Sigma)$ such that $u_a \le u_b$. Then the allowable control set U is given by

$$U = \{u \in C^\alpha(\Sigma) \mid u_a \le u \le u_b\}. \tag{4}$$

In the following, when the dependence of a function on the space-time variable $(x, t) \in Q$ or $(s, t) \in \Sigma$ is clear we will simply use a dot " $\cdot$" to denote the variable. For example, we will write $q(s, t, y, u)$ as $q(\cdot, y, u)$

C. Solutions to the state equation

The existence, uniqueness and *a priori* $C^{2,\alpha}(\overline{Q})$ norm of the solution to the state equation are all needed in the study of optimal conditions. The Venttsel problems of parabolic equations have been well studied [1] and [2] where the existence of classical and distributional solutions are obtained under very general structure conditions for a class of quasilinear equations and boundary conditions. Since our main focus here is the optimal control problem we will not quote the general existence results from [1] and [2] but will reformat the

Theorem to cover only our equations. Throughout this paper we assume

$$\sigma \leq 0. \tag{5}$$

Theorem II.1 *Let the following conditions hold: (a) The domain Ω has a C^3 boundary Γ; (b) $f \in C^\alpha(\overline{Q})$; (c) φ is a bounded smooth function. Then for each $u \in C^\alpha(\Sigma)$ the problem (2) has a solution $y \in C^{2,\alpha}(\overline{Q})$.*

The uniqueness of the classical solution is not given in [1] and [2] for the general problem. We will prove it for our problem by using the weak maximum principle and the arguments in [9].

Theorem II.2 *Under the assumptions of Theorem II.1 and (5) the solution $y \in C^{2,\alpha}(\overline{Q})$ to the problem (2) is unique.*

Proof: Suppose y_1 and y_2 are two solutions of (2). Then $y_3 = y_1 - y_2$ satisfies

$$\begin{aligned} D_t y_3 - \Delta y_3 &= 0 \text{ in } Q, \\ D_t y_3 - \Delta_\Gamma y_3 + \partial_\nu y_3 &= \sigma y_3, \text{ on } \Sigma, \\ y_3(x,0) &= 0 \text{ in } \Omega. \end{aligned}$$

By the weak maximum principle, (Theorem 6, Sec. 2 of [9]), the positive maximum of y_3 must be obtained at a point $P \in \Sigma$. Also, by Theorem 14, Sec. 2.2 of [9] $\partial_\nu y_3(P) > 0$ if P has the inside strong sphere property which is satisfied by our assumption $\Gamma \in C^3$. However at P we also have $D_t y_3(P) \geq 0$, $\Delta_\Gamma y_3(P) \leq 0$ and $\sigma y_3(P) \leq 0$ which gives a contradiction to the boundary condition. The same argument applies to the negative minimum of y_3. Therefore $y_3 \equiv 0$. □

To obtain the *a priori* bound for the $C^{2,\alpha}(\overline{Q})$ norm of the solution, the following weak maximum principle is the starting point. This result can be found in [20].

Theorem II.3 *Assume that there is a constant $M > 0$ such that $|\sigma| \leq M$ for all $(s,t) \in \Sigma$. Then the solution y to the problem (2) satisfies*

$$\|y\|_Q \leq C(\|y_0\|_\Omega + \|\varphi\|_\Sigma + \|f\|_Q) \tag{6}$$

where C depends only on T and M.

Proof: Let $v = e^{-(M+1)t}y$. We then have $\max|y| \leq e^{(M+1)T}\max|v|$. Assume v attains a positive maximum at (x_0,t_0) in $\overline{Q}$ then we have the following three cases:

Case I: $(x_0,t_0) \in Q$. Then

$$\begin{aligned} D_t v - \Delta v + (M+1)v &= e^{-(M+1)t}(D_t y - \Delta y) \\ &= e^{-(M+1)t} f. \end{aligned} \tag{7}$$

Since $\Delta v \leq 0$ and $D_t v \geq 0$ at (x_0,t_0), we have $v \leq e^{-(M+1)t_0} f$ and hence $\max v \leq \max|f|$.

Case II: $(x_0,t_0) \in \Sigma$. Then a similar calculation yields

$$D_t v - \Delta_\Gamma v + \partial_\nu v + (M+1)v - \sigma v = e^{-(M+1)t}\varphi. \tag{8}$$

Since $|\sigma| \leq M$, $\Delta_\Gamma v \leq 0$, $\partial_\nu v \geq 0$ and $D_t v \geq 0$ at (x_0,t_0), we have $v \leq e^{-(M+1)t_0}\varphi$ and hence $\max v \leq \max|\varphi|$.

Case III: $(x_0,t_0) \in \Omega$. Then $t_0 = 0$, $v = y$ so $\max v \leq \max|y_0|$.

In summary we have

$$\max y \leq e^{(M+1)T}(\max_\Omega |y_0| + \max_\Omega |\varphi| + \max_Q |f|).$$

If $\min y < 0$ then the same argument as above gives

$$-\min y \leq e^{(M+1)T}(\max_\Omega |y_0| + \max_\Omega |\varphi| + \max_Q |f|).$$

Thus the Theorem is proved. □

For $u \in U$ there exists a constant $M_1 > 0$ depending only on u_a and u_b such that $|\varphi| \leq M_1$ for all (s,t). In such a case (6) becomes

$$\|y\|_Q \leq C(M_1 + \|y_0\|_\Omega + \|f\|_Q). \tag{9}$$

We also need the *a priori* estimate for the $C^{2,\alpha}(\overline{Q})$ norm of the solution. We formulate the result to cover our simple problem only instead of giving the general result under complicated assumptions. Then we will briefly outline the idea of the proof instead of giving the detailed proof.

Theorem II.4 *Let the assumptions of Theorem II.1 and II.2 hold. Assume also that $|D^2\varphi| \leq M$ for a constant $M > 0$. Then there exists a constant C such that the solution $y \in C^{2,\alpha}(\overline{Q})$ to the problem (2) satisfies*

$$\|y\|_{C^{2,\alpha}(\overline{Q})} \leq C(M + \|y_0\|_{C^{2,\alpha}(\overline{\Omega})} + \|f\|_{C^\alpha(\overline{Q})}) \tag{10}$$

where C depends on n, diam(Ω), T and M.

Proof: Corollary 9 provides a bound for $\|y\|_Q$. Using this in Theorem 1.3 of [1] we obtain a bound for $\|Dy\|_{C^\alpha(\overline{Q})}$, in particular, for $\|\partial_\nu y\|_{C^\alpha(\overline{Q})}$. Then we rewrite the boundary condition in the form

$$D_t y - \Delta_\Gamma y - \sigma y_3 = \rho$$

where $\rho = -\partial_\nu y + \varphi(\cdot,0,u) \in C^\alpha(\Sigma)$ and σ is as before. This can be regarded as linear parabolic equation on the boundary Σ. By the interior estimate, Theorem 5 of Sec. 3.2 of [9], we then have

$$\|y\|_{C^{2,\alpha}(\Sigma)} \leq C_1(\|y_0\|_{C^{2,\alpha}(\Sigma)} + \|\rho\|_{C^\alpha(\Sigma)}).$$

Finally (10) follows from Theorem 6 of Sec. 3.2 of [9]. □

III. Differentiability

In order to derive the optimal condition we investigate the differentiability of the functionals involved in the problem and establish the expressions for the derivatives in this section. For this purpose we start with the *principal system* which is an initial-boundary value problem:

$$\begin{aligned} D_t y - \Delta y &= f \quad \text{in } Q, \\ D_t y - \Delta_\Gamma y + \partial_\nu y &= h, \quad \text{on } \Sigma, \\ y(x,0) &= y_0 \quad \text{in } \Omega, \end{aligned} \tag{11}$$

where $f \in C^\alpha(Q)$, $y_0 \in C^{2,\alpha}(\Omega)$ and $h \in C^\alpha(\Sigma)$ are given functions. We call the following system the *adjoint problem* of (11).

$$\begin{aligned} -D_t z - \Delta z &= g \quad \text{in } Q, \\ -D_t z - \Delta_\Gamma z + \partial_\nu z &= r, \quad \text{on } \Sigma, \\ z(x,T) &= z_T \quad \text{in } \Omega, \end{aligned} \tag{12}$$

where $g \in C^{\alpha}(Q)$, $z_T \in C^{2,\alpha}(\Omega)$ and $r \in C^{\alpha}(\Sigma)$ are given functions. For the pair of system (11) and (12) we have the following relation.

Theorem III.1 *Suppose that y is a solution of (11) and z is a solution of (12). Then the following formula holds*

$$\int_Q fz\ dxdt + \int_\Sigma yr\ dsdt = \int_Q yg\ dxdt + \int_\Sigma zh\ dsdt + \int_\Omega (y(x,T)z_T - y_0 z(x,0))\ dx - \int_\Gamma (y(x,T)z_T - y_0 z(x,0))\ ds \quad (13)$$

Proof: We multiply the differential equation in (11) by z and integrate both side over Q to get

$$\int_Q zD_t y\ dxdt - \int_Q z\Delta y\ dxdt = \int_Q fz\ dxdt. \quad (14)$$

From the integration by parts formula we have

$$\int_Q zD_t y\ dxdt = \int_\Omega \int_0^T zD_t y\ dtdx = \int_\Omega (y(x,T)z_T - y_0 z(x,0))\ dx - \int_Q yD_t z\ dxdt. \quad (15)$$

For the second term in (14) we have, by using Green's formula and the boundary condition in (2),

$$\begin{aligned}\int_Q z\Delta y\ dxdt &= \int_Q DyDz\ dxdt - \int_Q D(zDy)\ dxdt \\ &= \int_Q DyDz\ dxdt - \int_\Sigma z\partial_\nu y\ dsdt \\ &= \int_Q DyDz\ dxdt + \int_\Sigma z(D_t y - \Delta_\Gamma y)\ dsdt \\ &\quad - \int_\Sigma zh\ dsdt \end{aligned} \quad (16)$$

Applying integration by parts formula with respect to t and applying the boundary version of Green's identity (see Lemma 16.1 of [10]) with respect to $s \in \Gamma$ we have

$$\int_\Sigma z(D_t y - \Delta_\Gamma y)\ dsdt = \int_\Gamma (y(x,T)z_T - y_0 z(x,0))\ ds - \int_\Sigma y(D_t z + \Delta_\Gamma z)\ dsdt = \int_\Gamma (y(x,T)z_T - y_0 z(x,0))\ ds + \int_\Sigma yr\ dsdt - \int_\Sigma y\partial_\nu z\ dsdt.$$

The last equation follows from the boundary condition of problem (12). Then (16) becomes

$$\int_Q z\Delta y\ dxdt = \int_Q DyDz\ dxdt - \int_\Sigma zh\ dsdt + \int_\Sigma yr\ dsdt - \int_\Sigma y\partial_\nu z\ dsdt + \int_\Gamma (y(x,T)z_T - y_0 z(x,0))\ ds \quad (17)$$

Substituting (17) and (15) into (14) gives

$$\int_Q fz\ dxdt = \int_\Omega (y(x,T)z_T - y_0 z(x,0))\ dx - \int_\Gamma (y(x,T)z_T - y_0 z(x,0))\ ds - \int_\Sigma yr\ dsdt - \int_Q yD_t z - \int_Q DyDz\ dxdt + \int_\Sigma zh\ dsdt + \int_\Sigma y\partial_\nu z\ dsdt \quad (18)$$

On the other hand, by multiply the differential equation in (12) by y and integrating over Q we have

$$-\int_Q yD_t z\ dxdt - \int_Q y\Delta z\ dxdt = \int_Q yg\ dxdt \quad (19)$$

which is

$$-\int_Q yD_t z\ dxdt - \int_Q DyDz\ dxdt + \int_\Sigma y\partial_\nu z\ dsdt = \int_Q yg\ dxdt. \quad (20)$$

From this (13) follows. □

In the following, for convenience, we express the output y_u corresponding to the control u as the image of a mapping $G : C^{\alpha}(\Sigma) \to C^{2,\alpha}(Q)$ so that $y_u = G(u)$.

Theorem III.2 *The mapping $y = G(u)$ is twice Fréchet differentiable. If $G'(u) \in \mathcal{L}(C^{\alpha}(\Sigma), C^{2,\alpha}(Q))$ and $G''(u) \in \mathcal{L}(C^{\alpha}(\Sigma) \times C^{\alpha}(\Sigma), C^{2,\alpha}(Q))$ are the first and second order Fréchet derivative of G at u, then for each $v, v_1, v_2 \in C^{\alpha}(\Sigma)$ the function $z = \langle G'(u), v\rangle$ is the unique solution of the boundary value problem*

$$\begin{aligned} D_t z - \Delta z = 0 \quad &\text{in } Q, \qquad z(x,0) = 0 \quad \text{in } \Omega, \\ D_t z - \Delta_\Gamma z + \partial_\nu z = \sigma z + \frac{\partial \varphi}{\partial u}(\cdot, y, u)v \quad &\text{on } \Sigma \end{aligned} \quad (21)$$

and the function $z_{12} = \langle G''(u), (v_1, v_2)\rangle$ is the unique solution of the boundary value problem

$$\begin{aligned} D_t z - \Delta z = 0 \quad &\text{in } Q, \qquad z(x,0) = 0 \quad \text{in } \Omega, \\ D_t z - \Delta_\Gamma z + \partial_\nu z = \sigma z + \frac{\partial^2 \varphi}{\partial u^2}(\cdot, y, u)v_1 v_2 \quad &\text{on } \Sigma \end{aligned} \quad (22)$$

where $z_i = \langle G'(u), v_i\rangle$ for $i = 1, 2$.

Proof: We first prove that G is Gateaux-differentiable and calculate the G-derivative $dG(u)$. Let $v \in C^{\alpha}(\Sigma)$ and consider $y_\lambda = G(u + \lambda v)$ and $y = G(u)$. It follows that

$$z = \langle G'(u), v\rangle = \lim_{\lambda \to 0} \frac{w_\lambda}{\lambda}$$

where $w_\lambda = y_\lambda - y$ satisfies

$$\begin{aligned} &D_t w_\lambda - \Delta w_\lambda = 0 \quad \text{in } Q, \quad w_\lambda(x,0) = 0 \quad \text{in } \Omega. \\ &D_t w_\lambda - \Delta_\Gamma w_\lambda + \partial_\nu w_\lambda = \sigma w_\lambda + \varphi(\cdot, u + \lambda v) - \varphi(\cdot, u) \\ &\text{on } \Sigma. \end{aligned} \quad (23)$$

Dividing (23) by λ we can see that $z_\lambda = w_\lambda/\lambda$ satisfies

$$\begin{aligned} &D_t z_\lambda - \Delta z_\lambda = 0 \quad \text{in } Q, \quad z_\lambda(x,0) = 0 \quad \text{in } \Omega. \\ &D_t z_\lambda - \Delta_\Gamma z_\lambda + \partial_\nu z_\lambda = \sigma z_\lambda + \gamma_\lambda v \quad \text{on } \Sigma. \end{aligned} \quad (24)$$

ISBN: 978-1-61804-253-8

where

$$\gamma_\lambda = \int_0^1 \frac{\partial \varphi}{\partial u}(\cdot, u + \tau\lambda v)\, d\tau.$$

We can assume that λ is bounded, say $|\lambda| \leq 1$. Notice that $\|\gamma_\lambda v\|_{C^\alpha(\Sigma)}$ is also bounded and hence Theorem II.4 implies

$$\|z_\lambda\|_{C^{2,\alpha}(\overline{Q})} \leq C_3(\|z_\lambda\|_Q + 1)$$

for some constants C_3. Applying the maximum principle Theorem 6 we know that $\|z_\lambda\|_Q$ is bounded. In summary we

$$\|z_\lambda\|_{C^{2,\alpha}(\overline{Q})} \leq C_4 \tag{25}$$

for a constant C_4 independent of λ. This implies that, up to a subsequence, z_λ converges to a function z in $C^{2,\alpha}(\overline{Q})$ as $\lambda \to 0$ and

$$\lim_{\lambda \to 0} \gamma_\lambda = \frac{\partial \varphi}{\partial u}(\cdot, u).$$

By taking limit in (24) we can see that $z = \langle dG(u), v\rangle$ is the solution of (21).

The uniqueness of z is guaranteed by Theorem II.2.

Next we examine the continuity of dG. Notice that $dG(u) \in \mathcal{L}(C^\alpha(\Sigma), C^{2,\alpha}(\overline{Q}))$ and

$$\|dG(u)\| = \sup_{\|v\|=1} |\langle dG(u), v\rangle|_{C^{2,\alpha}(\overline{Q})}.$$

Therefore to prove the continuity of $dG(u)$ is to prove that as $\tilde{u} \to u$ in $C^\alpha(\Sigma)$

$$\begin{aligned} &\|dG(\tilde{u}) - dG(u)\| \\ &= \sup_{\|v\|=1} |\langle dG(\tilde{u}), v\rangle - \langle dG(u), v\rangle|_{C^{2,\alpha}(\overline{Q})} \to 0. \end{aligned}$$

For any $v \in C^\alpha(\Sigma)$ with $\|v\| = \|v\|_{C^{2,\alpha}(\overline{Q})} = 1$ consider $\tilde{z} = \langle dG(\tilde{u}), v\rangle$ and $z = \langle dG(u), v\rangle$. Then $w = \tilde{z} - z$ satisfies

$$\begin{aligned} &D_t w - \Delta w = 0 \quad \text{in } Q, \qquad w(x,0) = 0 \quad \text{in } \Omega. \\ &D_t w - \Delta_\Gamma w + \partial_\nu w = \sigma w + \frac{\partial \varphi}{\partial u}(\cdot, y, \tilde{u})v - \frac{\partial \varphi}{\partial u}(\cdot, y, u)v \\ &\qquad\qquad \text{on } \Sigma. \end{aligned} \tag{26}$$

All we need to show is that $w \to 0$ in $C^{2,\alpha}(\bar{Q})$ uniformly with respect to $\|v\| = 1$, as $\tilde{u} \to u$ in $C^\alpha(\Sigma)$. To this end we put

$$\eta = v\frac{\partial \varphi}{\partial u}(\cdot, \tilde{u}) - v\frac{\partial \varphi}{\partial u}(\cdot, u).$$

and

$$\xi = \int_0^1 \frac{\partial^2 \varphi}{\partial u^2}(\cdot, y, u + \tau(\tilde{u} - u))\, d\tau.$$

Then the right hand side of the boundary condition (26) can be written as

$$\sigma w + \xi v(\tilde{u} - u).$$

From the assumption on φ we know that $\xi v \in C^\alpha(\overline{Q})$ and hence

$$\|\eta\|_{C^\alpha(\Sigma)} \leq C_5 \|\tilde{u} - u\|_{C^\alpha(\Sigma)}.$$

By Theorem II.4 we then have

$$\|w\|_{C^{2,\alpha}(\overline{Q})} \leq C_6 \|\tilde{u} - u\|_{C^\alpha(\Sigma)} \to 0$$

which proves the continuity of $dG(u)$. Finally, since $G(u)$ is continuously Gateaux differentiable, we conclude that $G(u)$ is also Fréchet differentiable and that the Fréchet derivative $G'(u)$ is equal to $dG(u)$.

For the second order derivative we let $y_\lambda = G(u + \lambda v_2)$ and $z_\lambda = \langle G'(u + \lambda v_2), v_1\rangle$. We then have

$$z_{12} = \langle G''(u), (v_1, v_2)\rangle = \lim_{\lambda \to 0} \frac{1}{\lambda}(z_\lambda - z_1).$$

By using exactly the same argument above we can prove the existence of $G''(u)$ and derive the equation that z_{12} must satisfy. Since this is a lengthy but straight forward exercise we omit the details here. □

Now we are in the position to establish the differentiability and express the derivative of the objective functional $J(u)$.

Theorem III.3 *The functional J is twice Fréchet differentiable and for every $u, v, v_1, v_2 \in C^\alpha(\Sigma)$ and $y = G(u)$ we have*

$$\langle J'(u), v\rangle = \int_\Sigma \left[\frac{\partial q}{\partial u}(\cdot, y, u) - \frac{\partial \varphi}{\partial u}(\cdot, y, u)w\right] v\, dsdt \tag{27}$$

and

$$\begin{aligned} \langle J''(u), (v_1, v_2)\rangle &= \int_Q \frac{\partial^2 p}{\partial y^2}(\cdot, y) z_1 z_2\, dxdt \\ &+ \int_\Sigma \frac{\partial^2 q}{\partial y^2}(\cdot, y, u) z_1 z_2\, dsdt \\ &+ \int_\Sigma \frac{\partial^2 q}{\partial u \partial y}(\cdot, y, u)(z_1 v_2 + z_2 v_1)\, dsdt \\ &+ \int_\Sigma \left[\frac{\partial^2 q}{\partial u^2}(\cdot, y, u) - \frac{\partial^2 \varphi}{\partial u^2}(\cdot, y, u)w\right] v_1 v_2\, dsdt \end{aligned} \tag{28}$$

where $z_i = \langle G'(u), v_i\rangle$ for $i = 1, 2$ and w is the solution of

$$\begin{aligned} &-D_t w - \Delta w = \frac{\partial p}{\partial y}(\cdot, y) \quad \textit{in } Q, \qquad w(x,T) = 0 \quad \textit{in } \Omega. \\ &-D_t w - \Delta_\Gamma w + \partial_\nu w = \sigma w - \frac{\partial q}{\partial y}(\cdot, y, u) \quad \textit{on } \Sigma. \end{aligned} \tag{29}$$

Proof: Define a mapping $H : C^{2,\alpha}(Q) \times C^\alpha(\Sigma) \to \mathbb{R}$ by

$$H(y, u) = \int_Q p(x, t, y)\, dxdt + \int_\Sigma q(s, t, y, u)\, dsdt.$$

Obviously H is differentiable with respect to both y and u. Also, for every $\tilde{y}$ and $\tilde{u}$ we have

$$\begin{aligned} \langle \frac{\partial H}{\partial y}(y, u), \tilde{y}\rangle &= \int_Q \frac{\partial p}{\partial y}(x, t, y(x,t))\tilde{y}\, dxdt \\ &+ \int_\Sigma \frac{\partial q}{\partial y}(s, t, y(s,t), u(s,t))\tilde{y}\, dsdt \end{aligned}$$

and

$$\langle \frac{\partial H}{\partial u}(y, u), \tilde{u}\rangle = \int_\Sigma \frac{\partial q}{\partial u}(s, t, y(s,t), u(s,t))\tilde{u}\, dsdt.$$

Since $J(u) = H(G(u), u)$, by the chain rule we have

$$\begin{aligned} \langle J'(u), v\rangle &= \langle \frac{\partial H}{\partial y}(y, u)G'(u) + \frac{\partial H}{\partial u}(y, u), v\rangle \\ &= \langle \frac{\partial H}{\partial y}(y, u), G'(u)v\rangle + \langle \frac{\partial H}{\partial u}(y, u), v\rangle \end{aligned} \tag{30}$$

where $G'(u)v$ stands for the solution $z = \langle G'(u), v\rangle$ of (21) in Theorem III.2. Therefore

$$\langle J'(u), v\rangle = \int_Q \frac{\partial p}{\partial y}(\cdot, y)z \; dxdt + \int_\Sigma \frac{\partial q}{\partial y}(\cdot, y, u)z \; dsdt + \int_\Sigma \frac{\partial q}{\partial u}(\cdot, y, u)v \; dsdt \quad (31)$$

Now we set (21) as the principal system and treat (27) as its adjoint system. Let w be the solution of (27). Applying Theorem III.1 to z and w together with the information $z_0 = 0$, $w_T = 0$, $f = 0$, $g = \dfrac{\partial p}{\partial y}(\cdot, y)$, $h = \sigma z + \dfrac{\partial \varphi}{\partial u}(\cdot, y, u)v$ and $r = \sigma w - \dfrac{\partial q}{\partial y}(\cdot, y, u)$ we obtain

$$\int_\Sigma z\left(\sigma w - \frac{\partial q}{\partial y}(\cdot, y, u)\right) \; dsdt = \int_Q z\frac{\partial p}{\partial y}(\cdot, y) \; dxdt + \int_\Sigma w\left(\sigma z + \frac{\partial \varphi}{\partial u}(\cdot, y, u)v\right) \; dsdt$$

which is

$$\int_Q z\frac{\partial p}{\partial y}(\cdot, y) \; dxdt = -\int_\Sigma z\frac{\partial q}{\partial y}(\cdot, y, u)) \; dsdt - \int_\Sigma w\frac{\partial \varphi}{\partial u}(\cdot, y, u)v \; dsdt$$

A substitution of this into (27) gives

$$\langle J'(u), v\rangle = \int_\Sigma \left(\frac{\partial q}{\partial u}(\cdot, y, u) - \frac{\partial \varphi}{\partial u}(\cdot, y, u)w\right) v \; dsdt.$$

For the second order derivative we differentiate $\langle J'(u), v_1\rangle$ using the formula (30) to get

$$\begin{aligned}
\langle J''(u), (v_1, v_2)\rangle &= \langle \frac{\partial H}{\partial y}(y, u), z_{12}\rangle + \langle \frac{\partial^2 H}{\partial y^2}(y, u), z_1 z_2\rangle \\
&+ \langle \frac{\partial^2 H}{\partial u^2}(y, u), v_1 v_2\rangle + \langle \frac{\partial^2 H}{\partial u \partial y}(y, u), (z_1 v_2 + z_2 v_1\rangle \\
&= \int_Q \frac{\partial p}{\partial y}(\cdot, y)z_{12} \; dxdt + \int_\Sigma \frac{\partial q}{\partial y}(\cdot, y, u)z_{12} \; dsdt \\
&+ \int_Q \frac{\partial^2 p}{\partial y^2}(\cdot, y)z_1 z_2 \; dxdt \\
&+ \int_\Sigma \frac{\partial^2 q}{\partial y^2}(\cdot, y, u)z_1 z_2 \; dsdt + \int_\Sigma \frac{\partial^2 q}{\partial u^2}(\cdot, y, u)v_1 v_2 \; dsdt \\
&+ \int_\Sigma \frac{\partial^2 q}{\partial u \partial y}(\cdot, y, u)(z_1 v_1 + z_2 v_2) \; dsdt \quad (32)
\end{aligned}$$

Then we set (22) as the principal system and treat (29) as its adjoint system. Let z_{12} and w be the solutions of (22) and (29) respectively. Applying Theorem III.1 to z_{12} and w together with the information $z_0 = 0$, $w_T = 0$, $f = 0$, $g = \dfrac{\partial p}{\partial y}(\cdot, y)$, $r = \sigma w - \dfrac{\partial q}{\partial y}(\cdot, y, u)$ and $h = \sigma z + \dfrac{\partial^2 \varphi}{\partial u^2}(\cdot, y, u)v_1 v_2$, we obtain

$$\int_Q z\frac{\partial p}{\partial y}(\cdot, y) \; dxdt = -\int_\Sigma z\frac{\partial q}{\partial y}(\cdot, y, u)) \; dsdt - \int_\Sigma w\frac{\partial \varphi}{\partial u}(\cdot, y, u)v \; dsdt$$

A substitution of this into (32) produces the formula (28) for $\langle J''(u), (v_1, v_2)\rangle$. □

IV. Optimality Condition

When the optimization problem doesn't have state constraints but has only the constraint (4) on the control u, a first order necessary optimality condition for $\bar{u}$ to be a solution is that $\langle J'(\bar{u}), v\rangle \geq 0$ for all v chosen from the allowable control set U. In the case $\langle J'(\bar{u}), v\rangle = 0$ for some $v \in U$ then the second order necessary optimality condition is $\langle J''(\bar{u}), (v, v)\rangle \geq 0$ for those v. The following necessary optimality condition is a straight forward consequence of Theorem III.3.

Theorem IV.1 *The necessary condition for $\bar{u} \in U$ to be an optimal solution of $\inf J(u)$ is*

$$\int_\Sigma \left(\frac{\partial q}{\partial u}(s, t, \bar{y}, \bar{u}) - \frac{\partial \varphi}{\partial u}(s, t, \bar{y}, \bar{u})\bar{w}\right) v \; dsdt \geq 0 \quad (33)$$

for all $v \in U$, where the couple $(\bar{y}, \bar{w})$ is the solution of the following system

$$\begin{cases}
D_t y - \Delta y = f \text{ in } Q, \\
D_t y - \Delta_\Gamma y + \partial_\nu y = \sigma y + \varphi(\cdot, \bar{u}) \text{ on } \Sigma, \\
y(x, 0) = y_0 \quad \text{in } \Omega, \\
\\
-D_t w - \Delta w = \dfrac{\partial p}{\partial y}(\cdot, y) \quad \text{in } Q, \\
-D_t w - \Delta_\Gamma w + \partial_\nu w = \sigma w - \dfrac{\partial q}{\partial y}(\cdot, y, \bar{u}) \quad \text{on } \Sigma, \\
w(x, T) = 0 \quad \text{in } \Omega.
\end{cases} \quad (34)$$

In the case where the integral in (33) is equal to 0 for some v then the second order necessary optimality condition is

$$\begin{aligned}
&\int_Q \frac{\partial^2 p}{\partial y^2}(\cdot, y)z^2 \; dxdt \\
&+ \int_\Sigma \frac{\partial^2 q}{\partial y^2}(\cdot, y, u)z^2 \; dsdt \\
&+ 2\int_\Sigma \frac{\partial^2 q}{\partial u \partial y}(\cdot, y, u)zv \; dsdt \\
&+ \int_\Sigma \left[\frac{\partial^2 q}{\partial u^2}(\cdot, y, u) - \frac{\partial^2 \varphi}{\partial u^2}(\cdot, u)w\right] v^2 \; dsdt \geq 0 \quad (35)
\end{aligned}$$

for such v, where $z = \langle G'(\bar{u}), v\rangle$.

As a special case we have:

Corollary IV.2 *If the objective functional $J(u)$ is convex, then the necessary and sufficient condition for $\bar{u}$ to be an optimal solution is*

$$\frac{\partial q}{\partial u}(s, t, \bar{y}, \bar{u}) - \frac{\partial \varphi}{\partial u}(s, t, \bar{u})\bar{w} = 0 \quad (36)$$

where the couple $(\bar{y}, \bar{w})$ is the solution of (34).

Proof: In such a case (33) becomes a necessary and sufficient condition with the equality holds true for all v. This implies (36). □

For an application of Corollary IV.2 we consider the example when $\varphi(\cdot, u) = u$, $\sigma = 0$ and the objective function is given by

$$J_1(u) = \frac{1}{2}\int_Q (y_u - y_g)^2 \; dxdt + \frac{\beta}{2}\int_\Sigma u^2 \; dsdt$$

where $\beta > 0$ is a constant and y_g is a given reference temperature. The objective in this example is to minimize the difference between the actual temperature and a given reference temperature plus the cost of the control. To verify that $J(u)$ is also convex with respect to u we let y_1, y_2 and y_λ be the solution of the state equation (2) corresponding to the controls u_1, u_2 and $\lambda u_1 + (1-\lambda)u_2$ respectively. It is easy to see that $y_\lambda = \lambda y_1 + (1-\lambda)y_2$ from the equation (2) and hence the convexity of $J_1(u)$ follows. Therefore, in such a case, we can find the optimal solution precisely by solving a system of heat equations.

Theorem IV.3 *If $\varphi(\cdot, y, u) = u$ then the optimal solution $\bar{u}$ of the problem $\inf J_1(u)$ is given by*

$$\bar{u} = \frac{1}{\beta}\bar{w}$$

where the pair $(\bar{w}, \bar{y})$ is the solution of the following system

$$\begin{cases} D_t y - \Delta y = f \text{ in } Q, \quad D_t y - \Delta_\Gamma y + \partial_\nu y = \dfrac{1}{\beta} w \quad \text{on } \Sigma, \\ y(x,0) = y_0 \quad \text{in } \Omega, \\ \\ -D_t w - \Delta w = y - y_g \quad \text{in } Q, \\ -D_t w - \Delta_\Gamma w + \partial_\nu w = 0 \quad \text{on } \Sigma, \;\; w(x,T) = 0 \quad \text{in } \Omega. \end{cases}$$

Proof: In Theorem IV.2 we put $p = (y - y_g)^2/2$, $q = (\beta u^2)/2$, $\varphi = u$. It follows that $\dfrac{\partial p}{\partial y} = y - y_g$, $\dfrac{\partial q}{\partial y} = 0$, $\dfrac{\partial q}{\partial u} = \beta u$, $\dfrac{\partial \varphi}{\partial u} = 1$ and $\dfrac{\partial \varphi}{\partial y} = 0$. Then the sufficient and necessary condition in Theorem IV.2 becomes $\beta\bar{u} = \bar{w}$, as long as $(\bar{w}, \bar{y})$ is the solution of (37). Therefore $\bar{u} = \frac{1}{\beta}\bar{w}$ is the optimal solution. □

REFERENCES

[1] Apushkinskaya D. E. and Nazarov A. I., *A survey of results on nonlinear Venttsel problems*, Applications of Mathematics, 45, No. 1, 69-80, (2000)

[2] Apushkinskaya, D. E.; Nazarov, A. I. *Hlder estimates of solutions to initial-boundary value problems for parabolic equations of nondivergent form with Wentzel boundary condition*, Amer. Math. Soc. Transl. (2) 64, 1-13, (1995)

[3] Apushkinskaya, D. E.; Nazarov, A. I. The nonstationary Venttsel problem with quadratic growth with respect to the gradient, J. Math. Sciences, Vol. 80, No. 6, 21972207, (1996)

[4] Ben Tal A. and Zowe J., *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Programming Study, 19, 39-76, (1999)

[5] Bonnans J. and Casas E., *Contrôle de systèmes elliptiques semilineaires comportant des contraintes sur l'ètat*, In Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, H. Brezis and J. lions, eds., vol 8, 69-86, Londonman Scientific & Technical, New York, (1988)

[6] Carslaw, H. S. and Jaeger, J. C., *Conduction of heat in solids* , Oxford, Clarendon Press, (1959)

[7] Casas E. and Tröltzsch F., *Second-order necessary optimality conditions for some state-Constrained control problems of semilinear elliptic equations*, Appl Math Optim, 39, 211-227, (1999)

[8] Casas E. and Tröltzsch F., *Second-order sufficient optimality conditions for some state-Constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., Vol. 38, No. 5, pp. 13691391 (2000)

[9] Friedman A., *Partial differential equations of parabolic type*, Dover Publications, New York, (1992)

[10] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of the Second Order*, 2nd Edition, Springer Verlag, (1975)

[11] Krumbiegel K. and Rehberg J. *Second order sufficient optimality conditions for parabolic optimal control problems with pointwise state constraints*, SIAM J. Control Optim., Vol. 51, No. 1, 304331, (2013)

[12] Lions J.-L., *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, 1971

[13] Luo Y., *Necessary optimality conditions for some control problems of elliptic equations with Venttsel boundary conditions*, Appl. Math. Optim., 61, 337-351, (2001)

[14] Luo Y. and Trudinger N.S., *Linear second order elliptic equations with Venttsel boundary conditions*, Proc. Royal Society of Edinburgh, 118A, 193-207, (1991)

[15] Luo Y. *Quasilinear second order elliptic equations with elliptic Venttsel boundary conditions*, Nonlinear Anal. 16, 761-769, (1991)

[16] Luo Y., *The heat conduction in a medium enclosed by a thin shell of higher diffusivity*, Proc. 5th Colloquium on Diff. Equations, Bulgaria, (1994)

[17] Raymond J.-P. and Tröltzsch F.-J. *Second order sufficient optimality conditions for nonlinear parabolic optimal control problems with state constraints*, Discrete and Continuous Dynamical Systems, Vol. 6, No. 2, 431450, (2000)

[18] Venttsel A. D., *On boundary conditions for multidimensional diffusion processes*, Theor. Probab. Appl., 4, 164-177, (1959)

[19] Robinson S. M., *First order conditions for general nonlinear optimization*, SIAM J. Appl. Math., Vol. 30 No.4, 597-607, (1976)

[20] Zeng Y., *Linear parabolic equations with venttsel initial boundary conditions*, Bul. Aus. Math. Society, Vol. 50, No. 3, 465-479, (1994)

[21] Zowe L. and Kurcyusz S., *Regularity and stability for the mathematical programming problem in Babach spaces*, Appl. Math. Optim., 5, 49-62, (1979)

Activities of Neuron and Unicellular Organism for Positive Pulse Generation

Atsushi Fukasawa and Yumi Takizawa

***Abstract*—**Activity of a neuron and realization of self-systematization of a neural group are presented. A $p - n$ boundary is analyzed to form a depletion layer in electrolyte. Where, p and n stand for major charges carrying signals. Electro-physical modeling of a neuron is given by two depletion layers in a neuron. A neuron is proved to operate as an amplifier or as a pulse generator to transmit electric signals. Electric modeling of a neural group is given by mutual coupling among neurons. This coupling provides a neural group with synchronization to realize fast and reliable signal processing. The membrane model is evaluated lastly based on electric and physical knowledge and theories.

***Keywords*—** Activity of neuron, electro-physical modeling, amplifier and pulse generator, self-systematization of neural group.

I. Introduction

IT is assumed essentially important to clarify principle of operations in neural systems in brain, which are almost unknown still now on.

Divergence has been the main aspect for the study of neural systems depending on the difference in spices, organ, tissues, and so on. Locations of active parts and paths of signals transmissions have been clarified. But the most interested information is the principles of operations and organizations of neural systems to realize sophisticated capabilities of functions against external and internal stimuli.

In this study, a neuron is analyzed by new basis of biology with electro-physical modeling of a neuron. It was found that a neuron operates as an amplifier or a pulse generator to transmit signal information.[1][2]

Because of defect of knowledge on operational principle of a neuron, essential study of systematization of neural group has not been done.

In this paper, a neural group is also analyzed by new basis of telecommunication system knowledge. It was found that a neural group operates as a synchronized system with holding common time inside the system.

This work was supported in part by the joint research project with Dr. Masaji Abe, COE, Musasino Co., Ltd., and by the trans-disciplinary project by Pro. Hiroe Tsubaki, Vice Director-General, the institute of Statistical Mathematics, Japan.

Yumi Takizawa is with the Institute of Statistical Mathematics, Tachikawa, Tokyo, 190-8562 Japan (phone: 81-50-5533-8539, fax: 81-42-526-4332; e-mail: takizawa@ism.ac.jp).

Atsushi Fukasawa was with Chiba University, Chiba, Japan. He is now with Musasino Co., Ltd., Ota-ku, Tokyo, Japan (e-mail:fukasawafuji@yahoo.co.jp).

In this paper, failures are presented regarding to the membrane model based on electrical and physical requirements.

II. Dynamics of Electric Charges at a Boundary in Electrolyte

A. Formation of electrical zones and a depletion layer

When electric charges are injected into a zone in electrical medium, charge density at the zone becomes higher and the other zone remains lower. It is assumed that quantity of injected charges is little and velocity of charges is low in the medium. Special phenomena are induced at a boundary between two zones as shown in Fig. 1.

Phase 1

Injected p-charges diffuse to n-zone, and n-charges diffuse to p-zone by the force of gradient of density F_D.

Phase 2

Coulomb's force FC (force by potential gradient) appears between diffused p- and n-ions. Directions of forces F_D and F_C are opposite. When they are balanced, diffusion is ceased.

Phase 3

A pair of space charges appears at both sides of the boundary. Potential difference appears in the boundary. And electric charges are driven outside the boundary, and two zones and a depletion layer formed at the boundary.

B. Depth and Capacitance of a Depletion Layer

Special phenomena described above is analyzed theoretically by the Maxwell-Hertz equations.

$$\left.\begin{aligned} &\mathrm{rot}\, E + \frac{\partial B}{\partial t} = 0 \\ &\mathrm{rot}\, H - \frac{\partial D}{\partial t} = i \\ &\mathrm{div}\, D = \rho \\ &\mathrm{div}\, B = 0 \end{aligned}\right\} \qquad (1)$$

$$D = \varepsilon_e E, \quad B = \mu_e H$$

ISBN: 978-1-61804-253-8

Phase 1: Diffusion of charges by gradient of density F_D.

Phase 2: Balance of diffusion F_D and Coulomb's force F_C.

Phase 3: Cease of diffusion and formation of;
(a) p-zone and n-zone, and
(b) space charges and depletion layer with depth *d*.

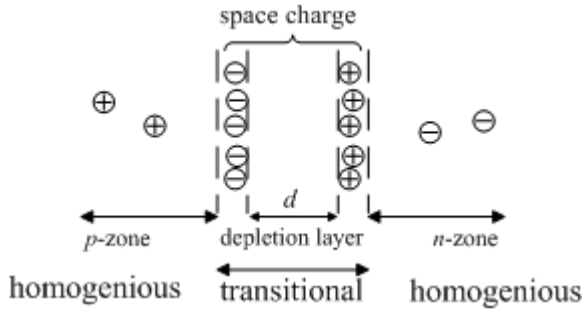


Fig. 1 Formation of zones and a depletion layer at a boundary.

where, *E*, *H* are electric and magnetic field strength, *D*, *B* are electric and magnetic flux density, and ε, μ are permittivity and permeability of medium respectively.

In cytoplasm, the followings are assumed;

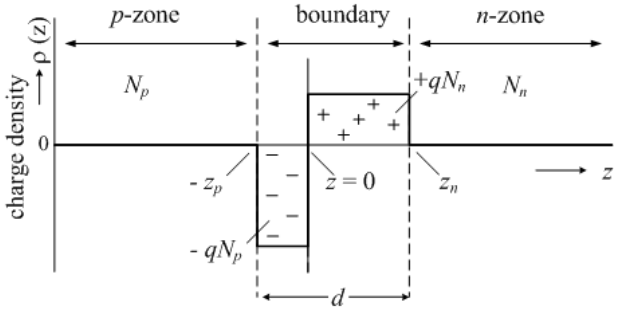
$$B = 0, \quad H = 0 \qquad (2),$$

$$\text{rot } E = 0 \qquad (3).$$

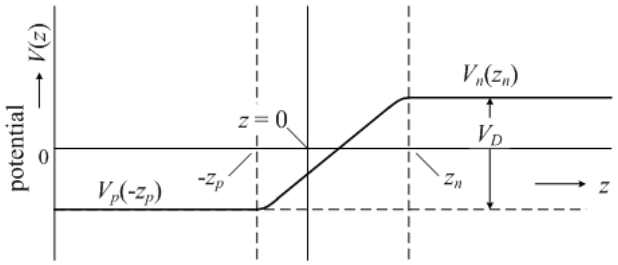
Eq.(1) reduces into Eq.(4).

$$\text{div} D = \rho \qquad (4),$$

Potential *V*(*z*) at a boundary is decided by true electric charge density $\rho(z)$ based on the Poison's equation.



(a) Distribution of true electric charge density ρ(*z*).



(b) Potentials *Vp*, *Vn*, and diffusion potential V_D.

Fig. 2 Distribution of true electric charge and diffusion potential of a boundary.

$$\frac{d^2V(z)}{dz^2} = -\frac{\rho(z)}{\varepsilon_e} \qquad (5)$$

where, *z* is the longitudinal axis of a neuron, ε_e is the permittivity of electrolyte solution.

True electric charge is defined as the charge unrestrained to any place. Then polarization charge at the membrane is removed from $\rho(z)$, because the polarization charge is restrained to the membrane in a neuron.

True electric charge density $\rho(z)$ is given by the followings, and is shown in Fig.2 (a).

$$\left.\begin{aligned} \rho(z) &= -q\,N_p \quad ; \quad -z_p \le z \le 0 \\ \rho(z) &= +q\,N_n \quad ; \quad 0 \le z \le z_n \end{aligned}\right\} \qquad (6)$$

where *Np*, *Nn* are true electric charge densities at *p*- and *n*-side of the boundary. *q* is elementary electric charge.

The electrical diffusion potential *VD* is defined as follows.

$$V_D = V_n(z_n) - V_p(-z_p)$$

$$= \frac{q}{2\varepsilon_e}(N_p z_p{}^2 + N_n z_n{}^2) \qquad (7)$$

The depth of depletion layer is given as,

$$d = z_p + z_n \quad = \left(\frac{2\varepsilon_e (N_p + N_n)}{q N_p N_n} V_D \right)^{\frac{1}{2}} \qquad (8)$$

Now, bias *VB* is assumed applied to a boundary. When *VB* is applied reversely to *n*-zone against *p*-zone, the depth of depletion layer d_B with reverse bias V_B is given as follows.

$$d_B = \left(\frac{2\varepsilon_e (N_p + N_n)}{q N_p N_n} (V_D + V_B) \right)^{\frac{1}{2}} \qquad (9)$$

The positive charge *Q* per unit area at the boundary (*n*-side) is given as follows.

$$Q = q N_n z_n = \left| -q N_p z_p \right| = \left(\frac{2\varepsilon_e q N_p N_n)}{N_p + N_n} (V_D + V_B) \right)^{\frac{1}{2}} \qquad (10)$$

The structure is assumed as an equivalent capacity.

$$c = \left| \frac{dQ}{dV} \right| = \left(\frac{\varepsilon_e}{2} \frac{q N_p N_n}{N_p + N_n} \frac{1}{V_D + V_B} \right)^{\frac{1}{2}} \qquad (11)$$

When *VB* is applied forwardly at the boundary, the capacity is given changing V_B to - V_B .

III. Electro-Physical Modelling of a Neuron

A. Depth and Capacity of Depletion Layers in a Neuron

(1) Whole aspects of a neuron

A neuron is exhibited as a three-port bio-electrical device with dendrite, central part, and axon. Here, the transmission line part is deleted in actual axon.

These ports are assigned as input, ground, and output ports. The ends of dendrite and axon are composed of multiple branches which are connected to previous and post neurons with synapses. Biochemical and electrical couplings are formed by synapses. A bio-electrical modeling is given in Fig. 3. An excitatory synapse is shown in the figure.

(2) Signal *p*-ion injection to a resting neuron

During a neuron is resting, inner potential is kept negative and uniform inside the neuron. When neurotransmitters are released from previous neurons and accepted by the neuron, *p*-charges of Na^+ are injected into the dendrite. Injected *p*-ions play as an excitatory signal into the neuron.

(3) Dynamics of signal *p*-ions at the first depletion layer

The first depletion layer is formed between the dendrite and the central parts.

The depth d_1 and the equivalent input capacity c_d are given as;

$$d_1 = \left\{ \frac{2\varepsilon_e (N_d + N_c)}{q N_d N_c} (V_{D1} - V_{B1}) \right\}^{\frac{1}{2}} \qquad (12)$$

$$c_d = \left(\frac{\varepsilon_e}{2} \frac{q N_d N_c}{N_d + N_c} \frac{1}{V_{D1} - V_{B1}} \right)^{\frac{1}{2}} \qquad (13)$$

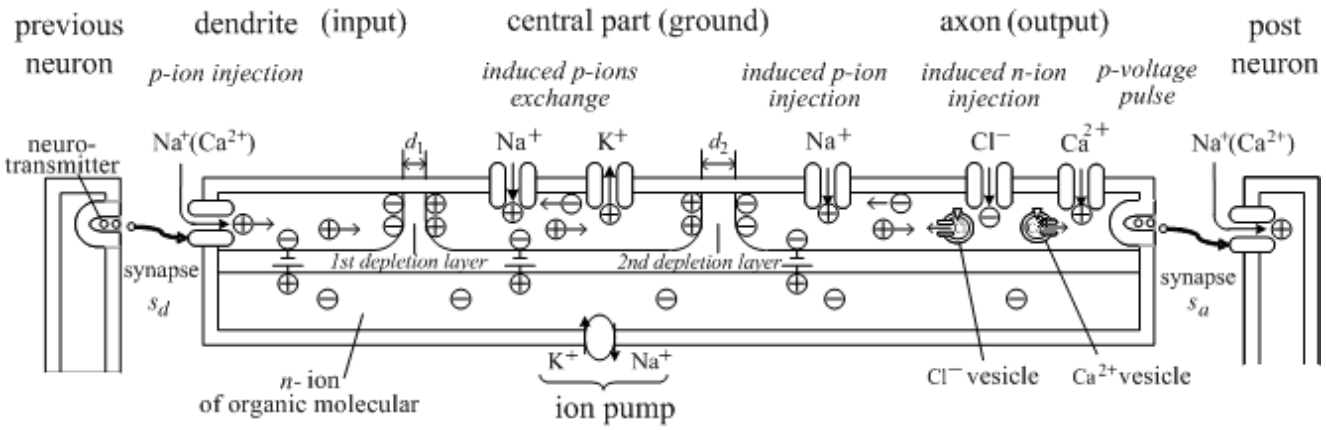


Fig. 3 Electro-physical modeling of an operating neuron. All branches of the dendrite and the axon of a neuron are gathered into each one port. Ion channels of the same kind are gathered in one in an area.

where, *Nd*, *Nc* are true electric charge density at the dendrite and the central part.

V_{D1} V_{B1} are diffusion potential and bias appeared forwardly at the first depletion layer.

The depth d_1 becomes narrower as injected *p*-ion at the dendrite and/or the forward-bias V_B increases. Then signal *p*-ions pass over the first depletion layer easily.

(4) *p*-ion injection to the axon

A part of signal *p*-ions reach the axon. Na^+ channels inject *p*-ions inside the neuron.

The potential in this zone is changed into positive, this area forms *p*-zone.

By Na^+ injection, charge allocation at the second depletion layer should be inverted as shown in Fig. 3. This configuration is defined by reverse diode with reverse bias voltage.

(5) Dynamics of signal *p*-ions at the second depletion layer

The second depletion layer is formed between the central part and the axon. The depth d_2 and the equivalent output capacity c_a are given as;

$$d_2 = \left\{ \frac{2\varepsilon_e \ (N_c + N_a)}{q \ N_c \ N_a} \ (V_{D2} + V_{B2}) \right\}^{\frac{1}{2}} \qquad (14)$$

$$c_a = \left(\frac{\varepsilon_e}{2} \frac{qN_cN_a}{N_c + N_a} \frac{1}{V_{D2} + V_{B2}} \right)^{\frac{1}{2}} \qquad (15)$$

where, *Nc* and *Na* are true electric charge density at the dendrite and the central part. V_{D2} V_{B2} are diffusion potential and bias appeared at the second depletion layer.

The depth d_2 becomes wider than depth d_1.

Signal *p*-ions pass over the second depletion layer by the force of thermal motion of ions.

(6) Dynamics of induced signal *n*-ions

When signal *p*-ions arrive at the axon, *n*-ions are injected into the axon by Cl^- channels.

n-ions move from the right to the left passing over the second and then the first depletion layers. The dynamics of *n*-ions from right to left is forward, and from left to right is reverse.

These *n*-ions play also as the signal together with signal *p*-ions. The *p*- and *n*-ions carry signals to the same direction with the principle of duality.

IV. Electrical Modeling of Activity of a Neuron

A. Formulation of Activity in a Neuron

Electrical modeling of an active neuron is shown in Fig. 4.

i_d is the current of *p*-ions injected in the dendrite, i_a is the current of sum of arrived *p*-ions and *n*-ions injected by Cl-channels at the axon. i_c is the current through resistance R_c of the central part to the outside of a neuron.

α is current multiplication factor and $\alpha \cdot i_d$ is equivalent current source for the axon.

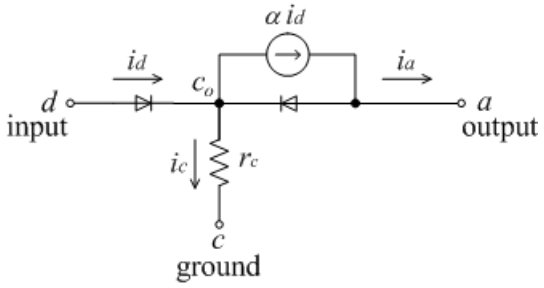


Fig. 4 Electrical modeling of activity of a neuron.

B. Characteristics as an Amplifier

Electrical modeling of an operating neuron is shown in Fig.7.

The points of d_0, a_0 are outside of membrane. c_0 is a virtual point taken in the central part. r_d and r_a are resistances of forward diode n_d and reverse diode n_a, r_c is the resistance at the central part to outside of a neuron. R_d and R_a are external resistances of synapses s_d and s_a..

$r_d << R_d$ and $r_a << R_a$. r_c is approximately zero.

The capacities C_d and C_a are caused by the first and second depletion layers respectively.

Input and output synapses s_d and s_a.are shown as forward diodes for excitatory synapses (*p*-ions). These synapses work as backward diodes for inhibitory synapses (*n*-ions).

Voltage amplification gain *G* is given as;

$$G = \frac{v_a}{v_d} = \frac{\dfrac{\alpha R_a}{r_d + r_c}}{1 - \dfrac{\alpha R_a}{r_d + r_c} \cdot \dfrac{r_c}{R_a}} = \frac{K}{1 - K\beta} \qquad (16)$$

$$K = \alpha \frac{R_a}{r_d + r_c} \qquad (17)$$

$$\beta = \frac{r_c}{R_a} \qquad (18)$$

where, v_d and v_a are input and output voltages of a neuron, *G*, *K*, β are closed loop gain, open loop gain, and inner feedback ratio of a neuron respectively. Oscillation condition is given by $K\beta \geq 1$.

In case that the axon has little Cl channels, $\alpha < 1$, $K\beta << 1$. Therefore a neuron operates as an amplifier with threshold for input signal with positive inner feedback.

C. Characteristics as a pulse generator

The neuron operates as an oscillator to generate pulses when the product of open loop gain *K* and feedback ration β exceeds 1.

This oscillator is composed by self injection without input trigger.

$$T_1 = C_d \frac{r_c R_a}{r_c + R_a} \quad (19)$$

$$T_2 = C_a R_a \quad (20),$$

where, $R_d + r_d >> r_c$, $r_a = \infty$

are assumed for simplified analysis.

The period of oscillation *T* is given as the total time length as following;

$$T = T_1 + T_2 = C_d \frac{r_c R_a}{r_c + R_a} + C_a R_a \quad (21).$$

D. Timing of output pulses

An oscillator operates in free running condition without external input. Timing of output pulse is adjusted in pull-in condition when external input i_d is added.

Output pulses v_a under free-running and pulled-in conditions are shown with dotted and solid lines in Fig. 6 respectively.

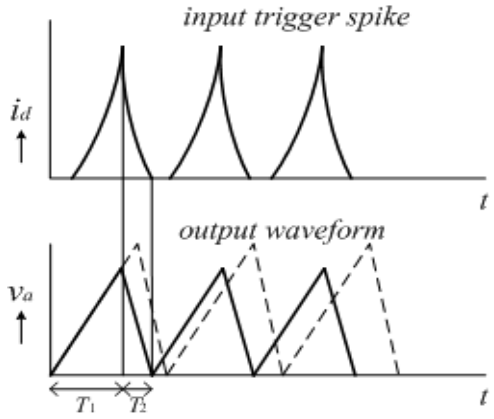


Fig. 6 Astable pulse generator by external injection.
Dotted line is an original waveform. Solid line is the waveform synchronized to input trigger pulse.

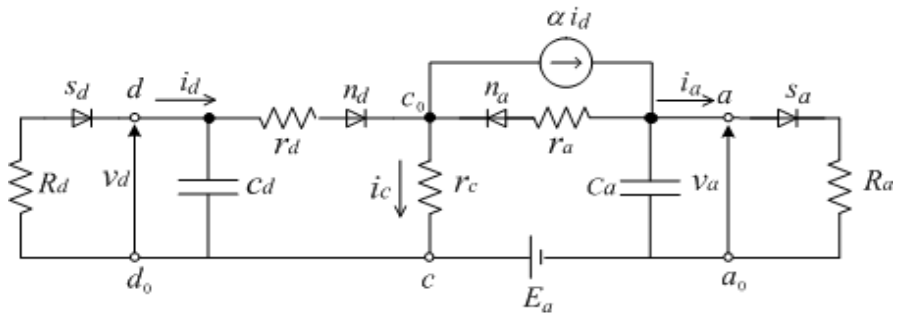


Fig. 5 Electrical modeling of an active neuron.

V. Self-Systematization by Mutual Pulse Injection among Neurons

A. Formation of a Neural Group

Actual formation of a neural group is given in Fig. 7. A small circle represents a neuron. Input and output signals of a neuron are at a branch of the dendrite and at a branch of the axon. A set of pair neurons is shown in Fig. 7 (a). Connection between two neurons is performed by arrows with dual directions. A system of four neurons is shown in Fig.7 (b).

B. System Synchronization

The timing of output pulse of an oscillator is adjusted by the other. When two oscillators are connected with each other, the timing is set at a certain timing between two. As number of oscillators increases, the variation of timings among neurons is reduced and system synchronization is established.

C. Synchronous Signal Processing

This formation enables system synchronization and synchronized signal processing simultaneously. Signal processing for multiple inputs and multiple outputs are available for dynamic processing including correlation, comparison, and detecting variations. This formation will be required for complex, reliable, and fast operation and signal processing [2]

VI. Energy Diagram of Electrical Charges in Active Neuron

Energy of *p*- and *n*-ions in a neuron are illustrated in Fig. 8. The energy of *p*- and *n*-ions are assumed with a small difference to Fermi level as shown in the figure.

Cl channels at the axon inject *n*-ions to left at the second depletion layer passing over a slope shown in the figure.

The dynamics of *p*- and *n*-ions is well informed by tracing the curve to right (*p*-ion) and to left (*n*-ion). The three port configuration is kept in spite with a slope at the axon (ref. [13,14]).

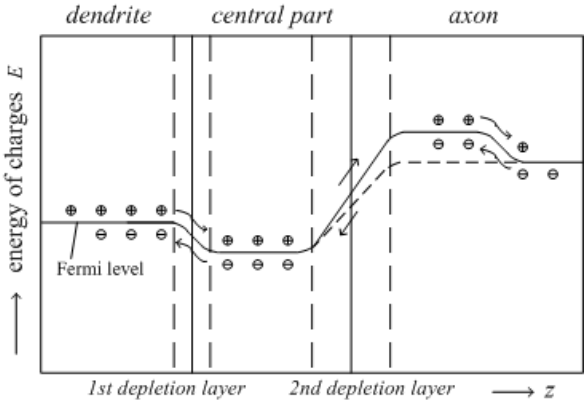
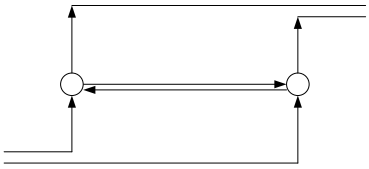
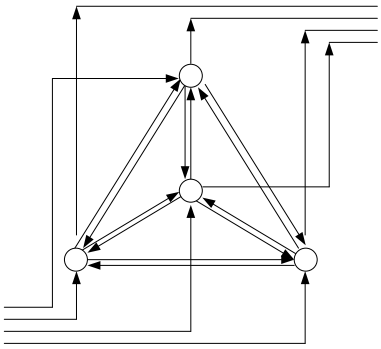


Fig. 8 Energy diagram of negative and positive ions with Cl^- channels at axon.



(a) A set of two neurons.



(b) A system by four neurons.

Fig. 7 Synchronization and signal processing by mutual injection.

ISBN: 978-1-61804-253-8

VII. Activity of Excitatory Cells of Paramecium and Neuron

Many kinds of cells generate recepter potential inside the cell for external stimulus. This phenomena is regarded as commnon response to thses kinds of cells. P*aramecium* is just examples of cells to study essential structure and method of operation for activity [8-10].

(a) *Paramecium caudatum*

Functional scheme of typical *paramecium* is shown in Fig.9 (a). This animal composes antierior (left), central part, and posterior (right). Macro- and Micro-neucleuses are at the central part, and cillia are at the surface. Recepter potentil is induced for stimulus at the front end. If the potential exceeds threshold, positive potential pulse is generated which drives motion of cillia in water of pond.

The body moves forward by motion of cillia for stimulus given at the rear end, and the body moves backward by reverse motion of cillia for stimulus given at the front end.

Increase of inner density of Ca^{2+} ion causes excitation of the cell, which is common to neuron, so *paramecium* is said as "swimming neuron".

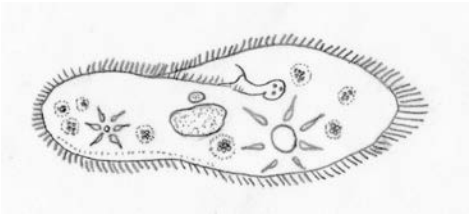


Fig. 9 Unicellular organisms, *Paramecium caudatum*, size: (0.2~0.3 mm) x 0.05 mm.

(b) Neuron

It is thought that anterior and posterior parts diverged to receptor and effector cells and organs. The central part diverged to neuron.

Output signal of a neuron is not transmitted electrically but chemically. The authors are informed by this study that this function originates in secretion of unicellular organism.

VIII. Conclusion

Activity of a neuron was presented for validation of the modelling referring unicellular organism with common behavior of excitatory cells.

Modelling was first given with electro-physical structure composed of three zones and two depletion layers formed in ectoplasm membrane for activity of a neuron.

An electrical modelling was then given for amplification and pulse generation. Positive pulse generation was shown by self and mutual injection. Stability of phase and period is realized by mutual injection to yield stable timing (clock) inside systems.

Validation of proposed modelling was given by comparison of activities by neuron and *paramecium* of unicellular organism.

Acknowledgment

The authors express their sincere gratitude for Prof. Hiro-aki Takeuchi, Department of Biology, Shizuoka University, Prof. Toshiharu Horie, Teikyo Heisei University, Associate prof., Kazuhiko Natori, central director of chemotherapy center, Toho University, prof. Alicia Gonzalo-Ruiz, Institute of Neuroscience of Castilla Leon, Spain, for their cooperation and advices to this study.

References

[1] Fukasawa A., Takizawa Y., Activity of a Neuron and Formulation of a Neural Group for Synchronization and Signal Processing, *Proc. of the Int. Conf. on Neurology*, pp.242-247, Kos, Greece, July 2012, "The Best Paper Prize of NEUROLOGY'12" awarded by WSEAS/NAUN.

[2] Fukasawa A., Takizawa Y., Activity of a Neuron and Formulation of a Neural Group based on Mutual Injection in keeping with system synchronization, *Proc. of International conference on Circuit, Systems, Control, Signals (CSCS'12)*, pp. 53-58, Barcelona, Spain, Oct. 17, 2012.

[3] Fukasawa A., Takizawa Y., Activity of a Neuron and Formulation of a Synchronous Neural System, *Proc. of the 15th International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems (MAMECTIS'13)*, pp. 66–73, Lemesos, Cyprus, Mar. 21-23, 2013.

[4] Fukasawa A., Takizawa Y., Activity of a Neuron and Self-Systematization of a Neural Group, *Proc. of International Conference on Biomedicine and Health engineering (BIHE'14)*, pp. 25-32, Tenerife, Spain, Jan. 10, 2014.

[5] Castsigeras E. Self-synchronization of networks with a strong kernel of integrate and fire excitatory neurons, *WSEAS Transactions on Mathematics*, Issue 7, Vol. 12, pp. 786 – 797, July 2013.

[6] Takizawa Y., Fukasawa A., Formulation of Topographical Mapping in Brain with a Synchronous Neural System, *Proceedings of the 15th International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems (MAMECTIS'13)*, pp. 60–65, Lemesos, Cyprus, Mar. 21-23, 2013.

[7] Fukasawa A. Takizawa Y., Activity of a Neuron and Formulation of a Neural Group for Synchronized Systems, *International Journal of Biology and Biomedical Engineering*, Issue 2, vol. 6, pp. 149-156, 2012.
[8] Kamada T., Some observations on potential difference across the ectoplasm membrane of Paramecium, *Journal of Experimental Biology*, vol. 11, pp.94-102, 1934.
[9] Naito Y., *Unicellular organisms and their Ethology*, (Japanese), University of Tokyo Press, Dec. 1990.
[10] Sakurai H., Takeuchi H., Mechanism for nervous system and behavior by endocrine disrupting chemicals – neuroethological and pharmacological analysis by using *Paramecium caudatum –, Proc. of the 14th annual meeting of JSEDR*, p.364, 2003.
[11] Takizawa Y., Fukasawa A., Formulation of a Neural System and Analysis of Topographical Mapping in Brain, *International Journal of Biology and Biomedical Engineering*, Issue 2, vol. 6, pp. 157-164, 2012.
[12] Fukasawa A., Takizawa Y., Activity of a Neuron brought by Electro-Physical Dynamics, *International Journal of Mathematical Models and Methods in Applied Sciences*, Issue 8, Volume 7, pp. 737-744, 2013.
[13] Shockley W., *Electrons and holes in semiconductors*, Fig. 4, pp. 112-113, D. Van, Nostrand, New York, 1950.
[14] Fukasawa A., Active circuit for antenna - Low noise semiconductor amplifier, *Master Thesis of Waseda Univ.* (Japanese), Mar. 1967.
[15] Takizawa Y., Rose G., Kawasaki M., Resolving Competing Theories for Control of the Jamming Avoidance Response: The Role of Amplitude Modulations in Electric Organ Discharge Decelerations, *Journal of Exp. Biol.* 202, pp. 1377-1386, 1999.
[16] Neher E, *Journal of Physiology*, pp. 193-214, 1988.
[17] Hille B., *Ion Channels of Excitable Membranes*, Sinauer Associates Inc., 2001.
[18] Fukasawa A., Takizawa Y., Electrical Measurement Scheme of Liquid Boundaries in Active Neuron, to be published on *Proc. of Int. Conf. on Health Science and Biomedical Systems (HSBS'14)*, Nov. 22, 2014.
[19] Fukasawa A., Takizawa Y., Activities of Neuron and Unicellular Organism as Excitatory Cells, to be published on *Proc. of Int. Conf. on Health Science and Biomedical Systems (HSBS'14)*, Nov. 22, 2014.
[20] Takizawa Y., Fukasawa A., Measurement of Boundary Position in Liquid Medium, to be published on *Proc. of Int. Conf. on MMCTSE'14*, Nov. 28, 2014.

Atsushi Fukasawa received the Master of Arts degree and the Ph.D. degree from Waseda University in 1967 and 1983. He joined Graduate School of Natural Science, Chiba University as a professor in 1997. He received the Award of the Agency of Science and Technology, Japan in 1982, and Ohm (publisher) Prize in 1994. He received Telecommunication System Technology Prize from the Foundation of Telecommunication Association, Japan in 2004. He is a senior member of the IEEE. He has been awarded for the Best Paper on NEUROLOGY'12.

Yumi Takizawa. Yumi Takizawa received the B.S. degree in Physics from Shinshu University in 1984, and the Ph.D. degree from the University of Tokyo in 1994. She joined the Institute of Statistical Mathematics (ISM) as an associate professor in 1995. She received the Prize on Telecommunication System Technology from the Foundation of Telecom Association, Japan in 2004. She has been engaged in neural systems in brain based on electro-physical and biological studies at the University of Virginia, USA and the ISM, Japan. She has been awarded for the Best Paper on NEUROLOGY'12.

Exploring the Applicability of bi-Helmholtz type nonlocal elasticity to the dynamical response of carbon nanotubes

C. Chr. Koutsoumaris, G. G. Vogiatzis, D. N. Theodorou and G. J. Tsamasphyros

Abstract**—**In the present study we investigate the problem of free vibrations of carbon nanotubes. Until now, all previous works on the dynamical response of single wall carbon nanotubes were based on Helmholtz-type nonlocal beam models. A study on bars has shown that the bi-Helmholtz model is the most appropriate one to fit Molecular Dynamics (MD) results. By considering results from various MD studies, we investigate the applicability of the bi-Helmholtz operator of nonlocal elasticity theory to approximate them. Moreover, we provide a summary of MD methods that have been applied to solve the free vibration problem.

Keywords**—**Nonlocal Elasticity, Beams, Carbon Nanotubes, eigenfrequencies, Molecular Dynamics.

I. Introduction

Since the discovery of carbon nanotubes (CNTs) at the beginning of the 1990s [1], extensive research related to them in the fields of chemistry, physics, materials science and electrical engineering has been reported. Mechanical behavior of CNTs, has been the subject of numerous studies [2 – 3].

This research has been co-financed by the European Union (European Social Fund—ESF) and Greek national funds through the Operational Program 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF). Project title: "THALIS-NTUA-Development of self healing composite materials and innovative techniques for structural health monitoring on aerospace applications"

C. Chr. Koutsoumaris is PhD student at the Division of Mechanics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens (NTUA), Zografou Campus 15773, Athens, Greece(phone: +30-210-772-4011; email: kkouts@mail.ntua.gr

G. G. Vogiatzis is PhD student at the Division of Materials Science and Engineering, School of Chemical Engineering, NTUA, Zografou Campus 15773, Athens, Greece (phone: +30-210-772-3216; email: gvog@chemeg.ntua.gr

D. N. Theodorou is a Professor at the Division of Materials Science and Engineering, School of Chemical Engineering, NTUA, Zografou Campus 15773, Athens, Greece (phone: +30-210-772-3157; email doros@central.ntua.gr

G. J. Tsamasphyros is a Professor Emeritus at the Division of Mechanics, School of Applied Mathematical and Physical Sciences, NTUA, Zografou Campus 15773, Athens, Greece (phone: +30-210-772-1297; email: tsamasph@central.ntua.gr

Since conducting controlled experiments at nanoscale is difficult and expensive, the research on the vibrational behavior of carbon nanotubes is directed towards molecular simulations and continuum theories.

Computer simulations allow the direct study of a simplified model and thus identification of whether discrepancies between theory and experiment are due to the simplifications of the model, to the approximations used in solving the theory, or to both. At first glance, Molecular Dynamics (MD) looks like a simplistic brute force attempt to literally reproduce what we believe is happening in the real world. Given a set of the initial positions and velocities, the equations of motion of all constituents of the system are integrated numerically. MD invokes a the description of the system in the classical limit, where the de Broglie thermal wavelength is much smaller than the mean nearest neighbor separation. The trajectory obtained from an MD run provides information necessary to calculate various time correlation functions, frequency spectra, diffusion coefficients, viscosity, and other dynamic and transport properties.

On the other hand, the applicability of classical continuum models at very small scales is questionable, since the material microstructure at small length scales, such as the lattice spacing between individual atoms, becomes increasingly important and the discrete structure of the material can no longer be homogenized into a continuum. Therefore, the modified continuum theories, such as nonlocal theory, may be an alternative route to consider the small scale effects in the studies of nanomaterials. The theory of nonlocal continuum mechanics and nonlocal elasticity was formally initiated by the papers of Eringen and Edelen [4 – 7]. Application of nonlocal continuum theory in nanomaterials was initially undertaken by Peddieson et al. [8], who applied the nonlocal elasticity to formulate a nonlocal version of Euler–Bernoulli beam model.

As far as the study of the free vibration problem for carbon nanotubes is concerned, there exists significant literature in the framework of non-local theory and MD simulations (the interested reader is referred to [9 – 12]). Most approaches on non local theory are based on the Helmholtz operator proposed by Eringen [4,7]. A study [13] on bars shows that the bi-Helmholtz model [14] is more appropriate to fit molecular dynamics results (Born-Karman). Considering this study, we investigate the bi-Helmholtz type nonlocal elasticity for the

free vibration problem of a single – wall carbon nanotube. Our analytical results are compared with MD obtained from our MD simulations and those available from the literature.

II. Governing Equations

A. General equations of Non Local Elasticity

For homogeneous and isotropic solids the linear elasticity theory is expressed by the following set of equations [4,7]

$$t_{kl,k} + \rho(f_l - \ddot{u}_l) = 0 \tag{1}$$

$$t_{kl}(\mathbf{x}) = \int_V K(|\mathbf{x}' - \mathbf{x}|, \tau)\sigma_{kl}(\mathbf{x}')d\upsilon'(\mathbf{x}') \tag{2}$$

$$\sigma_{kl}(\mathbf{x}') = \lambda e_{rr}(\mathbf{x}')\delta_{kl} + 2\mu e_{kl}(\mathbf{x}') \tag{3}$$

$$e_{kl}(\mathbf{x}') = \frac{1}{2}\left(\frac{\partial u_k(\mathbf{x}')}{\partial x_l'} + \frac{\partial u_l(\mathbf{x}')}{\partial x_k'}\right) \tag{4}$$

where t_{kl}, ρ, f_l and u_l are the components of the stress tensor, the mass density, the body force density vector components and the displacement vector components at a reference point $\mathbf{x}$ in the body at time *t*, respectively. Furthermore, $\sigma_{kl}(\mathbf{x}')$ is the classical stress tensor at $\mathbf{x}'$ which is related to the linear strain tensor $e_{kl}(\mathbf{x}')$ at any point $\mathbf{x}'$ in the body at time *t* via the Hooke law, with λ and μ being the Lamé coefficients. It can be readily observed that the only difference between (1) – (4) and the respective equations of classical elasticity is the expression of the stress tensor (2) which replaces Hooke's law (3). The volume integral in (2) is evaluated over the region *V* of the body.

Equation (2) expresses the contribution of other parts of the body to the stress at point $\mathbf{x}$ through the attenuation function (nonlocal modulus) $K(|\mathbf{x}' - \mathbf{x}|, \tau)$. From the structure of (2), we conclude that the attenuation function [6,14] has the dimensions of (length)$^{-3}$. Therefore, it should depend on a characteristic length ratio a/ℓ, where a is an internal characteristic length i.e. lattice parameter/bond length and ℓ is an external characteristic length i.e. crack length or wave length. Consequently, the expression of K in a more appropriate form is

$$K = K(|\mathbf{x}' - \mathbf{x}|, \tau),\ \tau = e_0 a / \ell \tag{5}$$

with e_0 being a dimensionless constant.

The nonlocal modulus exhibits the following properties:

i) When τ (or a) $\to 0$, *K* must reduce to the generalized Dirac function, so that classical elasticity limit is obtained at the limit of vanishing internal characteristic length.

$$\lim_{\tau \to 0} K(|\mathbf{x}' - \mathbf{x}|, \tau) = \delta(|\mathbf{x}' - \mathbf{x}|) \tag{6}$$

ii) It acquires its maximum at $\mathbf{x}' = \mathbf{x}$, attenuating with $|\mathbf{x}' - \mathbf{x}|$.

iii) If *K* is a Green's function of a linear differential operator i.e. if

$$L\left[K(|\mathbf{x}' - \mathbf{x}|, \tau)\right] = \delta(|\mathbf{x}' - \mathbf{x}|) \tag{7}$$

then by applying the operator *L* to eq.(2), we obtain:

$$Lt_{kl} = \sigma_{kl} \tag{8}$$

In this paper, we use the nonlocal modulus

$$K_{BH}(|x|) = \frac{1}{2}\frac{1}{c_1^2 - c_2^2}\left\{c_1 \exp(-|x|/c_1) - c_2 \exp(-|x|/c_2)\right\} \tag{9}$$

which is the Green's function of the following bi Helmholtz operator [7,14]

$$L^{BH} = \left(1 - c_1^2\frac{d^2}{dx^2}\right)\left(1 - c_2^2\frac{d^2}{dx^2}\right) = 1 - (c_1^2 + c_2^2)\frac{d^2}{dx^2} + c_1^2 c_2^2\frac{d^4}{dx^4}$$

$$= 1 - \varepsilon^2\frac{d^2}{dx^2} + \gamma^4\frac{d^4}{dx^4} \tag{10}$$

where $\varepsilon^2 = c_1^2 + c_2^2$ and $\gamma^4 = c_1^2 c_2^2$ c_1 and c_2 are given by the expressions:

$$c_1^2 = \frac{\varepsilon^2}{2}\left(1 + \sqrt{1 - 4\frac{\gamma^4}{\varepsilon^4}}\right)\quad c_2^2 = \frac{\varepsilon^2}{2}\left(1 - \sqrt{1 - 4\frac{\gamma^4}{\varepsilon^4}}\right)$$

where $0 \le \left(1 - 4\frac{\gamma^4}{\varepsilon^4}\right)$

If $4\gamma^4 = \varepsilon^4, c_1 = c_2$ are real and $\varepsilon = \sqrt{2}\gamma$. The kernel has the following form [14]:

$$K_{BH}(|x|, \gamma) = \frac{1}{2}\frac{1}{2\gamma^2}(\gamma + |x|)\exp(-|x|/\gamma)$$

$$K_{BH}(0, \gamma) = \frac{1}{4\gamma} \tag{11}$$

In the case that $\gamma = \frac{e_0 a}{\sqrt{2}}$ the modulus (9) can be written as:

$$K_{BH}(|x|, \gamma) = \frac{1}{2}\frac{1}{(e_0 a)^2}\left(\frac{e_0 a}{\sqrt{2}} + |x|\right)\exp\left(-|x|/\left(\frac{e_0 a}{\sqrt{2}}\right)\right) \tag{12}$$

with the corresponding operator taking the form:

$$L_1^{BH} = 1 - (e_0 a)^2\frac{d^2}{dx^2} + \frac{(e_0 a)^4}{4}\frac{d^4}{dx^4} \tag{13}$$

A variant of the above operator is for $\gamma = e_0 a$

$$L_2^{BH} = 1 - 2(e_0 a)^2\frac{d^2}{dx^2} + (e_0 a)^4\frac{d^4}{dx^4} \tag{14}$$

We also use the classical Helmholtz nonlocal modulus to make a comparison between the results obtained by each modulus.

The Helmholtz nonlocal modulus has the form:

$$K_H(|x|, \gamma) = \frac{1}{2e_0 a}\exp(-|x|/e_0 a) \tag{15}$$

which is the Green's function of the following operator [4,7]

$$L^H = 1 - (e_0 a)^2\frac{d^2}{dx^2} \tag{16}$$

B. Differential equation for Euler Bernoulli beam

The Euler Bernoulli beam theory (EBT) is based on the displacement field

$$u_1 = u(x,t) - z\frac{\partial w(x,t)}{\partial x} \qquad u_2 = 0 \qquad u_3 = w(x,t) \quad (17)$$

where (u,w) are the axial and transverse displacements of the point (x,0) on the middle plane of the beam. In the EBT the only nonzero strain is:

$$\varepsilon_{xx} = \frac{\partial u}{\partial x} - z\frac{\partial^2 w}{\partial x^2} \equiv \varepsilon_{xx}^0 + z\kappa \quad (18)$$

where ε_{xx}^0 is the extensional strain and κ is the curvature

The equations of motion are given by

$$\frac{\partial N}{\partial x} + f(x,t) = m_0\frac{\partial^2 u}{\partial t^2} \quad (19)$$

and

$$\frac{\partial^2 M}{\partial x^2} = m_0\frac{\partial^2 w}{\partial t^2} - m_2\frac{\partial^4 w}{\partial t^2\partial x^2} - q + \frac{\partial}{\partial x}\left(\hat{N}\frac{\partial w}{\partial x}\right) \quad (20)$$

Where $f(x,t)$, $q(x,t)$ are the axial force per unit length and transverse force per unit length, respectively. N is the axial force, M is the bending moment; they are defined as $N = \int_A \sigma_{xx} dA$, $M = \int_A z\sigma_{xx} dA$ with σ_{xx} being the classical – axial stress on the yz section in the x direction. The mass inertias m_0 and m_2 are defined by

$$m_0 = \int_A \rho dA = \rho A \text{ and } m_2 = \int_A \rho z^2 dA = \rho I$$

where I denotes the second moment of area about y axis and ρ denotes the mass density per unit length.

C. Differential equation for non local Euler Bernoulli beam

Applying the operator (10) in equation (8) and considering that $\sigma_{xx} = E\varepsilon_{xx}$,where E is the Young's modulus of the material, we obtain the following expressions:

$$L^H\tau_{xx} = \sigma_{xx}, L^{BH}t_{xx} = \sigma_{xx} \quad (21a,b)$$

or

$$\left(1 - \varepsilon^2\frac{\partial^2}{\partial x^2}\right)\tau_{xx} = E\varepsilon_{xx}, \left(1 - \varepsilon^2\frac{\partial^2}{\partial x^2} + \gamma^4\frac{\partial^4}{\partial x^4}\right)t_{xx} = E\varepsilon_{xx} \quad (22a,b)$$

where τ_{xx} and t_{xx} being the nonlocal stresses corresponding to Helmholtz $\left(L^H\right)$ and the bi-Helmholtz $\left(L^{BH}\right)$ operator.

The non local axial force N_{NL}^H, N_{NL}^{BH} and bending moment $M_{NL}^H\ M_{NL}^{BH}$ are defined as follows for each of the cases (21a,b):

i) $N_{NL}^H = \int_A \tau_{xx} dA \qquad M_{NL}^H = \int_A z\tau_{xx} dA \quad (23a,b)$

ii) $N_{NL}^{BH} = \int_A t_{xx} dA \qquad M_{NL}^{BH} = \int_A zt_{xx} dA \quad (24a,b)$

The process for constructing the differential equation refers to the operator L^{BH}.

Integrating the (21b) and using (24a) and (24b), we obtain:

$$\int_A L^{BH}t_{xx}dx = \int_A \sigma_{xx}dx \Rightarrow$$

$$N_{NL}^{BH} - \varepsilon^2\frac{\partial^2 N_{NL}^{BH}}{\partial x^2} + \gamma^4\frac{\partial^4 N_{NL}^{BH}}{\partial x^4} = EA\varepsilon_{xx}^o = N \quad (25)$$

$$\int_A L^{BH}zt_{xx}dx = \int_A z\sigma_{xx}dx \Rightarrow$$

$$M_{NL}^{BH} - \varepsilon^2\frac{\partial^2 M_{NL}^{BH}}{\partial x^2} + \gamma^4\frac{\partial^4 M_{NL}^{BH}}{\partial x^4} = EI\kappa = M \quad (26)$$

We consider the equations of motion:

$$\frac{\partial N_{NL}^{BH}}{\partial x} + f = m_0\frac{\partial^2 u}{\partial t^2} \text{ or } \frac{\partial N_{NL}^{BH}}{\partial x} = m_0\frac{\partial^2 u}{\partial t^2} - f \quad (27)$$

and

$$\frac{\partial^2 M_{NL}^{BH}}{\partial x^2} = m_0\frac{\partial^2 w}{\partial t^2} - m_2\frac{\partial^4 w}{\partial t^2\partial x^2} - q + \frac{\partial}{\partial x}\left(\hat{N}\frac{\partial w}{\partial x}\right) \quad (28)$$

Substituting for the first and third derivative of N from (27) into (25), we obtain

$$N_{NL}^{BH} = EA\frac{\partial u}{\partial x} + \varepsilon^2\left(m_0\frac{\partial^3 u}{\partial t^2\partial x} - \frac{\partial f}{\partial x}\right) - \gamma^4\frac{\partial^2}{\partial x^2}\left(m_0\frac{\partial^3 u}{\partial t^2\partial x} - \frac{\partial f}{\partial x}\right) \quad (29)$$

Substituting the second and fourth derivative of N_{NL}^{BH} from (29) into the equation of motion (27) we obtain:

$$\frac{\partial}{\partial x}\left(EA\frac{\partial u}{\partial x}\right) + f + \varepsilon^2\left(m_0\frac{\partial^4 u}{\partial t^2\partial x^2} - \frac{\partial^2 f}{\partial x^2}\right) - \gamma^4\left(m_0\frac{\partial^6 u}{\partial t^2\partial x^4} - \frac{\partial^4 f}{\partial x^4}\right) - m_0\frac{\partial^2 u}{\partial t^2} = 0 \quad (30)$$

Similarly, substituting the second and fourth derivative of M_{NL}^{BH} from (28) into (26) we obtain:

$$M_{NL}^{BH} = -EI\frac{\partial^2 w}{\partial x^2} + \varepsilon^2\left(m_0\frac{\partial^2 w}{\partial t^2} - m_2\frac{\partial^4 w}{\partial t^2\partial x^2} - q + \frac{\partial}{\partial x}\left(\hat{N}\frac{\partial w}{\partial x}\right)\right) - \gamma^4\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2} - m_2\frac{\partial^6 w}{\partial t^2\partial x^4} - \frac{\partial^2 q}{\partial x^2} + \frac{\partial^3}{\partial x^3}\left(\hat{N}\frac{\partial w}{\partial x}\right)\right) \quad (31)$$

Substituting M_{NL}^{BH} from (31) into (28) we obtain:

$$\frac{\partial^2}{\partial x^2}\left(-EI\frac{\partial^2 w}{\partial x^2}\right) + \varepsilon^2\frac{\partial^2}{\partial x^2}\left(m_0\frac{\partial^2 w}{\partial t^2} - m_2\frac{\partial^4 w}{\partial t^2\partial x^2} - q + \frac{\partial}{\partial x}\left(\hat{N}\frac{\partial w}{\partial x}\right)\right) - \gamma^4\frac{\partial^2}{\partial x^2}\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2} - m_2\frac{\partial^6 w}{\partial t^2\partial x^4} - \frac{\partial^2 q}{\partial x^2} + \frac{\partial^3}{\partial x^3}\left(\hat{N}\frac{\partial w}{\partial x}\right)\right) + q - \frac{\partial}{\partial x}\left(\hat{N}\frac{\partial w}{\partial x}\right) = m_0\frac{\partial^2 w}{\partial t^2} - m_2\frac{\partial^4 w}{\partial t^2\partial x^2} \quad (32)$$

In the case of pure bending *($m_2=0$)*and neglecting the axial force $\hat{N}$, we take:

$$M_{NL}^{BH} = -EI\frac{\partial^2 w}{\partial x^2} + \varepsilon^2\left(m_0\frac{\partial^2 w}{\partial t^2} - q\right) - \gamma^4\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2} - \frac{\partial^2 q}{\partial x^2}\right) \tag{33}$$

And the equation (32) takes the form:

$$\frac{\partial^2}{\partial x^2}\left(-EI\frac{\partial^2 w}{\partial x^2}\right) - \left(L_{NL}^{BH}\right)\left(m_0\frac{\partial^2 w}{\partial t^2} - q\right) = 0 \tag{34}$$

The same equation arises from the principle of virtual displacements as shown below:

$$\int_0^T\int_0^L\left\{\left[m_0\left(\frac{\partial u}{\partial t}\frac{\partial \delta u}{\partial t}\right) + m_2\left(\frac{\partial^2 w}{\partial x\partial t}\frac{\partial^2 \delta w}{\partial x\partial t}\right) + m_0\left(\frac{\partial w}{\partial t}\frac{\partial \delta w}{\partial t}\right)\right] - N_{NL}^{BH}\left(\frac{\partial \delta u}{\partial x}\right) + M_{NL}^{BH}\left(\frac{\partial^2 \delta w}{\partial x^2}\right) + \left[f(x,t)\delta u + q(x,t)\delta w\right]\right\}dxdt = 0 \tag{35}$$

Taking into account the (29) and (31) neglected m_2 [pure bending *($m_2=0$)*]and axial force $\hat{N}$, we have:

$$\int_0^T\left\{\int_0^L\left[m_0\left(\frac{\partial u}{\partial t}\frac{\partial \delta u}{\partial t}\right) + m_2\left(\frac{\partial^2 w}{\partial x\partial t}\frac{\partial^2 \delta w}{\partial x\partial t}\right) + m_0\left(\frac{\partial w}{\partial t}\frac{\partial \delta w}{\partial t}\right)\right]dx - \int_0^L\left[EA\frac{\partial u}{\partial x} + \varepsilon^2\left(m_0\frac{\partial^3 u}{\partial t^2\partial x} - \frac{\partial f}{\partial x}\right) - \gamma^4\left(m_0\frac{\partial^5 u}{\partial t^2\partial x^3} - \frac{\partial^3 f}{\partial x^3}\right)\right]\left(\frac{\partial \delta u}{\partial x}\right)dx + \int_0^L\left[-EI\frac{\partial^2 w}{\partial x^2} + \varepsilon^2\left(m_0\frac{\partial^2 w}{\partial t^2} - q\right) - \gamma^4\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2} - \frac{\partial^2 q}{\partial x^2}\right)\right]\left(\frac{\partial^2 \delta w}{\partial x^2}\right)dx + \int_0^L\left[f(x,t)\delta u + q(x,t)\delta w\right]dx\right\}dt = 0 \tag{36}$$

It can be verified that the Euler-Langrage equations associated with the variational statement in (36) are indeed the same as (30) and (34). Indicatively the bending equation *($m_2=0$)* is:

$$-\frac{\partial}{\partial t}\left(m_0\frac{\partial w}{\partial t}\right) - \frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2 w}{\partial x^2}\right) + \varepsilon^2\frac{\partial^2}{\partial x^2}\left(m_0\frac{\partial^2 w}{\partial t^2} - q\right) - \gamma^4\frac{\partial^2}{\partial x^2}\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2} - \frac{\partial^2 q}{\partial x^2}\right) + q(x,t) = 0 \tag{37}$$

In addition the natural boundary conditions at *$x= 0,L$* can be obtained:

$$\hat{Q} = -\frac{\partial}{\partial x}\left(EI\frac{\partial^2 w}{\partial x^2}\right) + \varepsilon^2\frac{\partial}{\partial x}\left(m_0\frac{\partial^2 w}{\partial t^2} - q\right) - \gamma^4\frac{\partial}{\partial x}\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2} - \frac{\partial^2 q}{\partial x^2}\right) \tag{38}$$

$$\hat{M} = -EI\frac{\partial^2 w}{\partial x^2} + \varepsilon^2\left(m_0\frac{\partial^2 w}{\partial t^2} - q\right) - \gamma^4\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2} - \frac{\partial^2 q}{\partial x^2}\right) \tag{39}$$

III. Free Vibtration Problems

In this section we study the free vibration problem for three different boundary conditions. We assume constant material and geometrical properties. The governing equation is obtained form (37).

For the free vibration we suppose that *$q(x,t)=0$*. Equation (37) takes the form:

$$-EI\frac{\partial^4 w}{\partial x^4} + \varepsilon^2\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2}\right) + \gamma^4\left(m_0\frac{\partial^6 w}{\partial t^2\partial x^4}\right) - \frac{\partial}{\partial t}\left(m_0\frac{\partial w}{\partial t}\right) = 0 \tag{40}$$

and the generalized forces become:

$$\hat{Q} = -EI\frac{\partial^3 w}{\partial x^3} + \varepsilon^2\left(m_0\frac{\partial^3 w}{\partial t^2\partial x}\right) - \gamma^4\left(m_0\frac{\partial^5 w}{\partial t^2\partial x^3}\right) \tag{41}$$

$$\hat{M} = -EI\frac{\partial^2 w}{\partial x^2} + \varepsilon^2\left(m_0\frac{\partial^2 w}{\partial t^2}\right) - \gamma^4\left(m_0\frac{\partial^4 w}{\partial t^2\partial x^2}\right) \tag{42}$$

In order to calculate the eigenfrequencies, we search for periodic solutions of the form $w(x,t) = \varphi(x)e^{i\omega t}$ where $\varphi(x)$ is the mode shape and **ω** is the natural frequency. After straightforward calculations, we obtain the following equations:

$$\left(EI - \gamma^4\omega^2 m_0\right)\frac{d^4\varphi}{dx^4} + \left(m_0\varepsilon^2\omega^2\right)\frac{d^2\varphi}{dx^2} - m_0\omega^2\varphi = 0 \tag{43}$$

$$\hat{Q}(x) = -EI\frac{d^3\varphi}{dx^3} - \varepsilon^2\omega^2\left(m_0\frac{d\varphi}{dx}\right) + \omega^2\gamma^4\left(m_0\frac{d^3\varphi}{dx^3}\right) \tag{44}$$

$$\hat{M}(x) = -EI\frac{d^2\varphi}{dx^2} - \omega^2\varepsilon^2\left(m_0\varphi\right) + \omega^2\gamma^4\left(m_0\frac{d^2\varphi}{dx^2}\right) \tag{45}$$

The general solution of (44) is:

$$\varphi(x) = c_1\sin(ax) + c_2\cos(ax) + c_3\sinh(\beta x) + c_4\cosh(\beta x) \tag{46}$$

where:

$$a^2 = \frac{1}{2\left(EI - \gamma^4\omega^2 m_0\right)}\left(m_0\varepsilon^2\omega^2 + \sqrt{4\left(EI - \gamma^4\omega^2 m_0\right)m_0\omega^2}\right)$$

$$\beta^2 = \frac{1}{2\left(EI - \gamma^4\omega^2 m_0\right)}\left(-m_0\varepsilon^2\omega^2 + \sqrt{4\left(EI - \gamma^4\omega^2 m_0\right)m_0\omega^2}\right)$$

A. Cantilever beam

For this problem the boundary conditions are:

$$\varphi(0) = 0,\ \varphi'(0) = 0 \quad \hat{M}(L) = 0, \hat{Q}(L) = 0 \tag{47}$$

Use of the boundary conditions leads to the condition:

$$c_1 + c_4 = 0, \quad c_1 a + c_4\beta = 0$$

$$\left(\left(EI - \gamma^4\omega^2 m_0\right)a^2 - \varepsilon^2 m_0\omega^2\right)\left(c_1\sin aL + c_2\cos aL\right) -$$

$$\left(\left(EI - \gamma^4\omega^2 m_0\right)\beta^2 + \varepsilon^2 m_0\omega^2\right)\left(c_3\sinh\beta L + c_4\cosh\beta L\right) = 0$$

$$a\left(\left(EI-\gamma^4\omega^2 m_0\right)a^2-\varepsilon^2 m_0\omega^2\right)\left(c_1\cos aL-c_2\sin aL\right)-$$
$$\beta\left(\left(EI-\gamma^4\omega^2 m_0\right)\beta^2+\varepsilon^2 m_0\omega^2\right)\left(c_3\cosh\beta L+c_4\sinh\beta L\right)=0 \quad (48)$$

In order to calculate the eigenfrequencies ω_n, $n=1,2,...N$ of a cantilever beam we have to set the determinant of the coefficient matrix in (48) to zero.

B. Simply supported beam

For this problem the boundary conditions are:

$$\varphi(0)=0,\ \varphi(L)=0,\ \ \hat{M}(0)=0,\hat{M}(L)=0$$

In order to calculate the eigenfrequencies ω_n, $n=1,2,...N$ of simply supported beam we follow the same procedure as that in the case of the cantilever beam.

C. Clamped clamped beam

For this problem the boundary conditions are

$$\varphi(0)=0,\ \varphi(L)=0,\ \varphi(0)=0,\ \varphi'(L)=0$$

Respectively for the eigenfrequencies ω_n, $n=1,2,...N$ of clamped-clamped beam we follow the same procedure as that of the cantilever beam and simply supported beam.

IV. Molecular Dynamics Simulations

A. Summary

Cornwell and Wille [15] were the first to use Molecular Dynamics (MD) simulations to investigate the elastic behavior of open-ended, free-standing, single wall carbon nanotubes. The MD simulations were carried out using the Tersoff-Brenner potential [16,17] and a standard Verlet algorithm for the integration of the equations of motion. The desired temperature was achieved by velocity rescaling. The tubes' response to axial compression was examined and typical failure modes as well as stress-strain curves for a number of tube radii were shown. The authors have calculated the Young's modulus of the tubes and developed a simple formula which approximated this quantity over a wide range of tube radii. Elastic properties have generally been found to be broadly consistent with the in-plane properties of graphite, but strengths have proved harder to assess, with simulation results consistently predicting higher values than have been observed experimentally.

Later, Elliott et al.[18] have used classical MD simulations to demonstrate that single wall carbon nanotube bundles collapse under hydrostatic pressure. In that study, a standard generic macromolecular force field, DREIDING [19] was used, which has been previously parameterized for carbon nanotubes [20]. MD simulations under constant number of particles, stress, and temperature were carried out on hexagonally close-packed bundles of ideal single wall carbon nanotubes (SWCNTs) with pseudoperiodic boundary conditions. The system stresses and temperature (298 K) were regulated using the Berendsen alogirthm, and the simulation time step was 1 fs. At pressures below 2 GPa, (10,10) SWCNTs showed only thermal motion of the atoms, with no apparent systematic deviation from the circular cross section. Above a critical pressure, the SWCNTs spontaneously collapse to form ribbons of oval cross section. The collapse pressures obtained as a function of nanotube diameter were in agreement with experimental data presented for small diameter nanotubes. It was found that the collapse pressure to be independent of the nanotube chirality.

Later, Duan et al.[11] have used MD simulations to parameterize Timoshenko beam theory models. Natural frequencies for a (5,5) armchair SWCNT with various length-to-diameter ratios under clamped – clamped and clamped – free boundary conditions were investigated. The interatomic interactions were described by the condensed-phased optimized molecular potential for atomistic simulation studies (COMPASS) force field [21], which has been proven to be applicable for describing the mechanical properties of carbon materials. The natural frequencies for the first four flexural modes of a (5,5) armchair SWCNT with various length-to-diameter ratios were estimated by MD simulations at room temperature. Those frequencies served as a basis for calibrating non-local elasticity theories. The boundary conditions employed were clamped-clamped or clamped – free. It was found that upon increasing the length-to-diameter ratio, the frequencies decrease. Furthermore, the frequencies under clamped-clamped boundary conditions were larger than those from clamped-free boundary condition.

A comprehensive MD study based on the COMPASS force field and continuum analysis was carried out by Cao et al.[22] to investigate the fundamental frequency shift of deformed clamped-clamped SWCNTs under axial loadings, bending and torsion. The results obtained by the beam model or the cylindrical shell model were found in good agreement with those obtained from MD simulations, provided that the Young's modulus and wall thickness are carefully selected. Yao and Lordi [23] performed MD simulations using the universal force field (UFF) to determine the Young's modulus of various clamped-free SWCNTs from their thermal vibration frequencies by using the frequency equations based on the Euler beam theory.

Zhang et al.[12] have performed MD simulations using the second-generation REBO potential for the bonding interactions, while fully neglecting the non-bonded interactions. Assuming that the temperature exerts no effect on the vibration frequency but it only increases the amplitude of vibrations, these authors performed MD simulations at a fixed temperature of 1 K. After an initial equilibration at this temperature, all atoms except those on the fixed end were allowed to vibrate. Then the vibration frequencies were computed by using the Fast Fourier Transform (FFT) of the velocity autocorrelation functions of the atoms.

More recently, Chang [24] has conducted MD simulations to investigate the vibrational behavior of the carbon nanotubes. Initially randomly assigned thermal vibrations of atoms were used to simulate the effect of the temperature. Since the vibration behavior of a free vibrating structure can be decomposed as a linear combination of all the resonant frequencies and their corresponding mode shapes, the resonant information could be analyzed from a free vibrating carbon nanotube under the influence of an initial thermal perturbation. All simulations were performed at room temperature, using a velocity rescaling method and the empirical Tersoff many-body potential [25], which is commonly adopted in CNT molecular simulation studies to provide a quick estimation and significant insight into the thermo-mechanical behavior.

B. Equations

In the microcanonical, or *NVE* ensemble, the system is evolved in time, by integrating the Newton's equation of motion:

$$\frac{d\mathbf{r}_i}{dt} = \nabla_{\mathbf{p}_i} H\,, \qquad \frac{d\mathbf{p}_i}{dt} = -\nabla_{\mathbf{r}_i} H \qquad (49a,b)$$

under constant number of particles, *N*, volume, *V*, and energy, *E*. The positions and momenta of the atoms are denoted by $\mathbf{r}_i$ and $\mathbf{p}_i$ respectively. In a classical description, H corresponds to the Hamiltonian of the system under investigation. This ensemble corresponds to an adiabatic process, where no heat exchange takes place. However, an exchange of potential and kinetic energy takes place, with total energy being conserved. Particle trajectories must stay on the appropriate constant-energy hypersurface in phase space, otherwise correct ensemble averages will not be generated.

For every timestep, each particle position, $\mathbf{r}_i$, must be integrated with a symplectic method, like the Verlet integrator: This method is a direct solution of the second-order equations above. The method is based on positions $\mathbf{r}_i(t)$, accelerations, $\mathbf{a}(t)$, and the positions $\mathbf{r}(t-\delta t)$ from the previous step. The timestep, δt should be sufficiently small to ensure energy conservation. The equation for advancing the positions reads as follows:

$$\mathbf{r}_i(t+\delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\delta t) + \delta t^2 \mathbf{a}(t) \qquad (50)$$

Along an MD trajectory, several correlations functions can be accumulated, like the velocity autocorrelation function defined as:

$$\mathcal{A}_i = \langle \mathbf{v}_i(t+\Delta t)\,\mathbf{v}_i(t) \rangle \qquad (51)$$

whose Fourier transform allows the estimations of the vibrational frequencies of the system under consideration.

V. Results And Discussion

A. Results

Numerical results are presented using the properties of carbon nanotubes, taking as reference the work of Zhang et al. [12], who studied the behavior of nanotubes utilizing the two most common approaches that in the literature about the elastic constants and the thickness of the nanotube.

The former of these arises from the fact that nanotubes are manufactured by graphene sheets which form the cylindrical structure, inheriting all its properties (thickness, elastic constants, etc.). The latter approach [26] comes from the initial observation that the intrinsic symmetry of a graphite sheet is hexagonal and the elastic properties of a two-dimensional hexagonal structure are isotropic, thus, it can be approximated by a uniform shell with only two elastic parameters : the flexural rigidity *D* and the in plane stiffness *C*. Considering small strains and equating the potential energy of the MD with the strain energy of shells we may calculate the parameters *C* and *D* and the corresponding Young modulus *E* and thickness *t* [26]. E=5.5 TPa & t= 0.066nm

In this work we exploited the results of the Zhang's et al. [12] MD simulations, as they coincide satisfactorily to our results obtained from MD simulations of carbon nanotubes with fixed end boundary conditions (cantilever). As far as the case of the transverse vibrations is concerned, the agreement is favorable, while as far as the longitudinal vibrations are concerned, the results obtained are identical.

In Figure 1 we present the spectrum of the transverse natural frequencies, as that was obtained by Fourier transform of the velocity autocorrelation functions, obtained from an MD trajectory under constant temperature, T = 300 K. The (5,5) SWCNT simulated had a length of 10 nm (9.1 nm of freely vibrating cantilever).

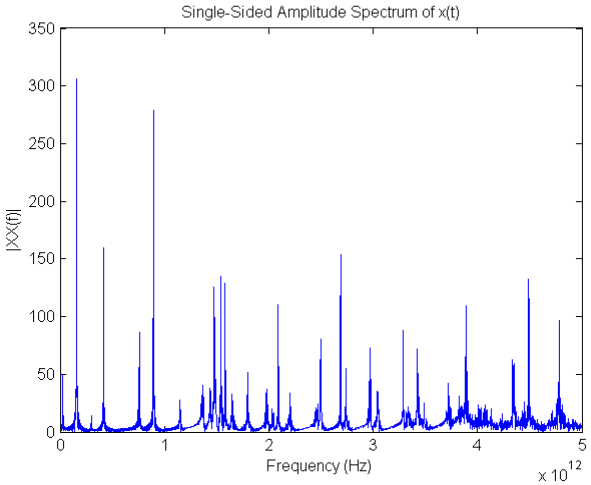


Figure 1. Amplitude spectrum along *xx'* direction

Zhang et al. have employed the classical elasticity theory in order to estimate the Young modulus and density of the SWCNT from the longitudinal vibration frequencies, avoiding defining the thickness of the SWCNT. Particularly interesting is the value of the density in the case where the thickness of the CNT is 0.066 nm, i.e. the nanotube has the characteristics of a shell structural element. The density is 11930 kg / m^3, which is almost five times larger than the density of carbon. However, it should be noted that the results obtained by following this scaling procedure, exhibit better agreement with those obtained from MD.

Our calculations have been based on the following values of the relative parameters [12] [(5,5) SWNT]:

Nanotube diameter $d = 2r = 0.68 \times 10^{-9}$ m ,

bond length $a = 0.142 \times 10^{-9}$ m

Beam Length $L = 13.69 \times d = 9.28 \times 10^{-9}$ m

a) thickness $t = 0.34 \times 10^{-9}$ m and $E = 0.845$TPa ,

$\rho = 2315 \text{kg/m}^3$ $I = 5.21 \times 10^{-38}\,\text{m}^4$ $A = 7.242 \times 10^{-19}\,\text{m}^2$

b) thickness $t = 0.066 \times 10^{-9}$ m and $E = 4.352$TPa ,

$\rho = 11930 \text{kg/m}^3$ $I = 8.15 \times 10^{-39}\,\text{m}^4$ $A = 1.46 \times 10^{-19}\,\text{m}^2$

Going to the results, in Figures (2),(3) and (4) we present the Ω_1 and Ω_3 dimensionless eigenfrequencies of the cantilever beam carbon nanotube (5,5), normalized to the corresponding MD eigenfrequency [12]. In Figures (5) and (6) we present the dimensionless Ω_1 and Ω_3 eigenfrequencies of the clamped-clamped beam carbon nanotube (5,5), normalized to the corresponding MD eigenfrequency [12].

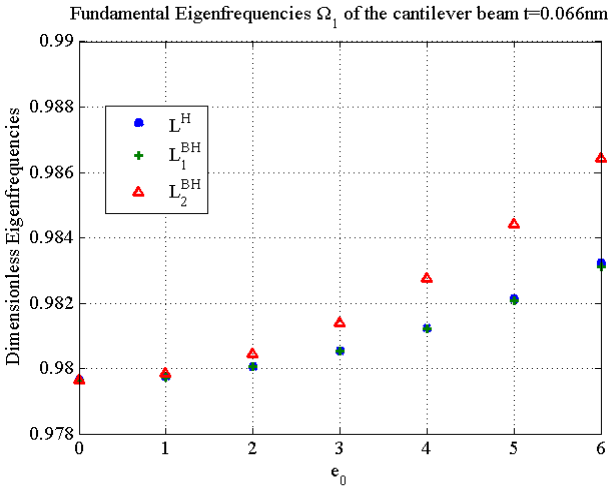


Figure 2. The Ω_1 eigenfrequency of cantilever beam Vs nonlocal parameter e_0 for the case of L^H, L_1^{BH}, L_2^{BH}.

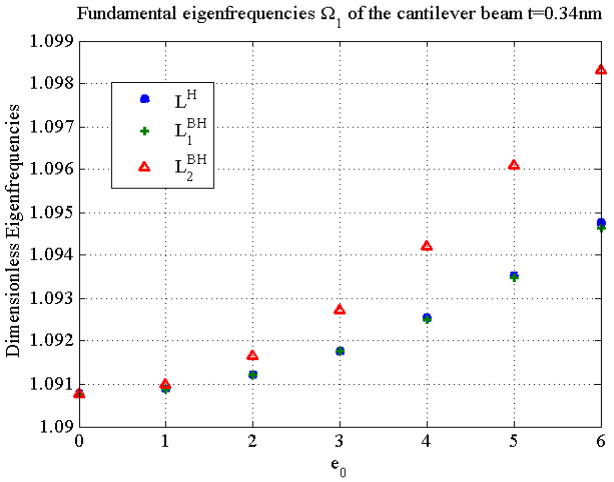


Figure 3. The Ω_1 eigenfrequency of cantilever beam Vs nonlocal parameter e_0 for the case of L^H, L_1^{BH}, L_2^{BH}.

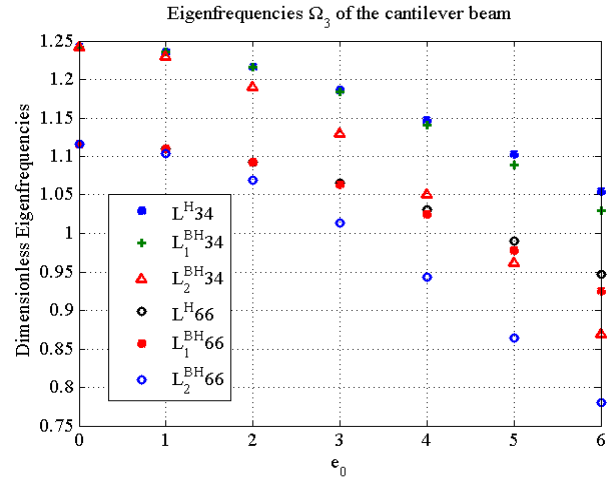


Figure 4. The Ω_3 eigenfrequency of cantilever beam Vs nonlocal parameter e_0 for the case of L^H, L_1^{BH}, L_2^{BH}. [The number (34 or 66) after the L and before the index is the thickness of nanotube (0.34nm or 0.066nm respectively)]

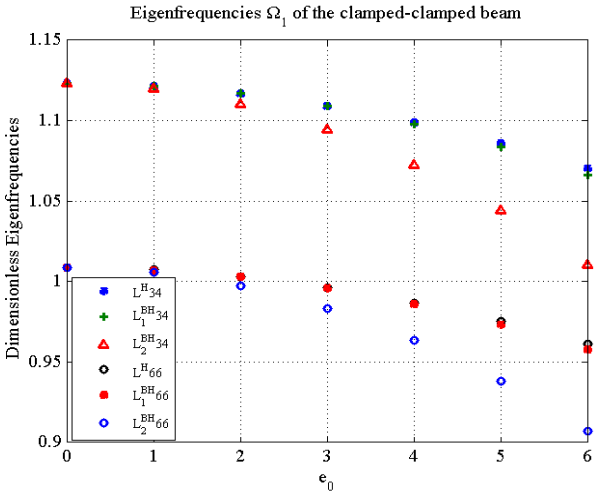


Figure 5 The Ω_1 eigenfrequency of the clamped-clamped beam Vs nonlocal parameter e_0 for the case of L^H, L_1^{BH}, L_2^{BH}

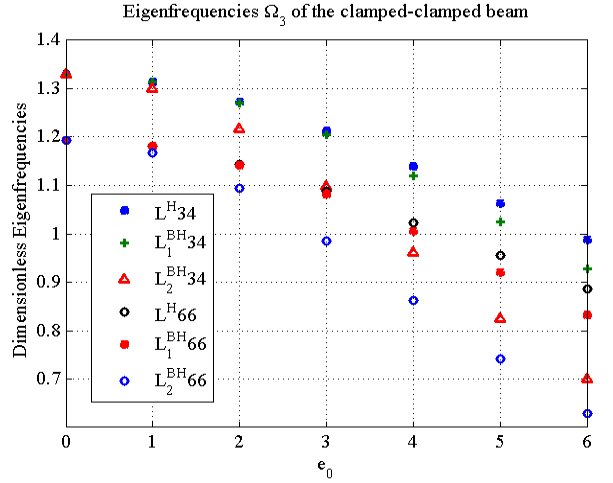


Figure 6. The Ω_3 eigenfrequency of the clamped-clamped beam Vs nonlocal parameter e_0 for the case of L^H, L_1^{BH}, L_2^{BH}

B. Discussion

Based on the above diagrams the following remarks and conclusions arise:

In Figure 2 we can see that in the case of *t=0.34* nm the classical $(e_0 a = 0)$ eigenfrequency (Ω_1) of the cantilever beam is approximately 9% greater than that of MD simulation. From Figure 3 we can see that in the case of *t=0.066* nm the classical $(e_0 a = 0)$ eigenfrequency (Ω_1) is approximately 2% less than that of MD simulation. Also, both L^H and the L^{BH} operator produce higher values of fundamental eigenfrequency with increasing e_0 in both cases of the thickness.

In Figure 4 we can see the Ω_3 eigenfrequency of the cantilever beam. It is evident that the better fit to the results of the MD simulation obtained for thickness of *t=0.066* nm. The use of nonlocal operators L^H, L^{BH} produce lower values of the eigenfrequency with increasing e_0 in both cases of the

ISBN: 978-1-61804-253-8

thickness. The L_2^{BH} operator exhibits a better fit than the other for the same value of e_0.

In Figures 5 and 6 we can see the fundamental eigenfrequency (Ω_1) and Ω_3 eigenfrequency of the clamped-clamped beam. It is evident that the better fit to the results of the MD simulation is obtained for a thickness of *t=0.066* nm. Again the use of nonlocal operators L^H, L^{BH} produce lower values of eigenfrequency with increasing e_0 in both cases of the thickness. And in this case, the L_2^{BH} operator has better fit than the other for the same value of e_0.

VI. Summary and Conclusions

The solution of the Euler–Bernoulli equations, in their classical formulation, assuming that CNTs can be approximated with shell features, describes better the results obtained from MD simulations. In both case studies, cantilever and clamped-clamped beam, the transverse frequencies are closer to those obtained from MD simulation.

The use of nonlocal formulation of the dynamic equations of motion, provides a better way to fit the MD simulations than the local, classical form. In particular, the use of the bi-Helmholtz operator L_2^{BH} gives better results than the Helmholtz operator L^H and bi-Helmholtz operator L_1^{BH} for the same range of parameter values in all cases (thickness, boundary conditions). Thus by using operator L_2^{BH} we obtain the same results but for lower values of the parameter e_0.

Finally, the estimation of the elastic constants directly from MD simulations would be of paramount importance. That would allow us to fully parameterize the analytical local and nonlocal models, given a forcefield describing the atomistic interactions of the CNT. On the one hand, the natural frequencies of CNTs can be obtained from MD trajectories, studying the velocity autocorrelation functions. On the other hand, a stress (or strain) fluctuation formalism can be invoked in order to predict the elastic stiffness (or elastic compliance) tensor, respectively. As already pointed out, scaling arguments for estimating model parameters, as that proposed by Zhang et al. lead to unrealistic values of the material properties (e.g. density).

References

[1] S. Iijima, "Helical microtubules of graphitic carbon", Nature 354, p. 56, 1991.

[2] A. Pantano, D.M. Parks, M.C. Boyce, "Mechanics of deformation of single- and multi- wall carbon nanotubes" , J. Mech. and phys. Sol. , vol 52, p. 789-821, 2004.

[3] V.M. Harik "Mechanics of carbon nanotubes: applicability of the continuum-beam models", Comp. Mat. Science, p 328-342, 2001.

[4] A.C. Eringen, "On differential equations of nonlocal elasticity and solutions of screw dislocation and surface waves", J. Appl. Phys., 54, pp.4703 – 4709, 1983

[5] A.C. Eringen and D. G. B. Edelen, "On nonlocal elasticity", Int. J. Engng. Sci. 10, pp. 233 – 248, 1972

[6] A.C. Eringen, "Theory of nonlocal elasticity and some applications", Res. Mechan. 21, pp.313 – 342, 1987.

[7] A.C. Eringen, *Nonlocal Continuum Field Theories*, Springer – Verlag, New York , 2002

[8] Peddieson, J., Buchanan, R.G., McNitt, P.R., "Application of nonlocal continuum models to nanotechnology", Int. J. Engng. Sci. vol 41,p. 305 – 312, 2002.

[9] J. N Reddy, "Nonlocal theories for bending, buckling and vibration of beams", International Journal of Engineering Science, vol. 45, pp. 288-307, June 2007.

[10] J. N Reddy and S.D. Pang,"Nonlocal continuum theories of beams for the analysis of carbon nanotubes", Journal of Applied Physics, vol. 103, pp. 023511, 2008.

[11] W.H. Duan, C.M. Wang and Y.Y. Zhang, "Calibation of nonlocal scaling effect for free vubration of carbon nanotubes by molecular dynamics", Journal of applied physics, vol. 101, 024305, 2007.

[12] Y.Y. Zhang, C.M. Wang an V.B.C. Tan, "Assessment of Timoshenko Beam models for vibrational behavior of single walled carbon nanotubes using molecular dynamics", Advances in Applied Mathematics and Mechanics, vol. 1, no.1, pp. 89-106, 2009.

[13] D.A. Fafalis, S.P. Filopoulos and G.J. Tsamasphyros, "On the capability of generalized continuum theories to capture dispersion characteristics at the atomic scale", European Journal of Mechanics A/Solids, vol. 36 pp 25-37, 2012.

[14] Lazar, M., Maugin, A.G., Aifantis, C. E., "On a theory of nonlocal elasticity of bi- Helmholtz type and some applications", Int. J. Solids Struct. 43, pp. 1404 – 1421, 2005.

[15] C.F. Cornwell, L.T. Wille, "Elastic properties of single-walled carbon nanotubes in compression", Sol. State. Comm, vol. 101 pp. 555 – 558, 1997.

[16] J. Tersoff, "Empirical Interatomic Potential for Carbon, with Application to Amorphous Carbon", Phys. Rev. Lett, vol 61, pp. 2879-2882 1988.

[17] D.W. Brenner, "Empirical potenrial for hydrocarbons for use in simulatinh the chemical vapor deposition of diamond films", Phys. Rev. B, vol. 42, p. 9458, 1990.

[18] J.A. Elliott, J.K. Sandler, A.H. Windle, Young, R.J. and Shaffer, M.S.P. "The collapse of single-wall carbon nanotubes is diameter-dependent", Phys. Rev. Lett., vol. 92, 095501,(2004).

[19] S.L. Mayo, B.D. Olafson, W.A. Goddard, "DREIDING: A Generic Force Field Molecular Simulations", J. Phys. Chem, vol. 94, p. 8897, 1990.

[20] M. Hanfald, H. Beister, K. Syassen, "Graphite under pressure: Equation of state and first-order Raman modes", Phys. Rev. B, vol. 39, p.12598, 1989.

[21] H. Sun, "COMPASS: An Ab Initio Forcefield Optimized for Condensed-Phase Application-Overview with Details on Alkane and Benzene Compounds," J. Phys. Chem, vol. B102, pp. 7338-7364, 1998.

[22] X.G. Gao, X. Chen, J.W. Kysar, "Strain sensing of carbon nanotubes: Numericla analysis of the vibrational frequency of deformed single wall carbon nanotubes", ", Phys. Rev. B, vol. 72, p.195412, 2005.

[23] N. Yao, V. Lordi, "Young's modulus of single-walled carbon nanotubes", J. Appl. Phys, vol. 84, p. 1939, 1998

[24] I-LingChang, " Molecular dynamics investigation of carbon nanotube resonance", Modeling and Simulation in Materials Science and Engineering, vol. 21, pp. 045011, 2013.

[25] J. Tersoff, " New empirical model for the structural properties of silicon", Phys. Rev. Lett, vol 56, p. 632 1986.

[26] B.I. Yakobson, C.J. Brabec, J. Bernholc, " Nanomechanics of carbon Tubes: Instabilities beyond linear response", Phys. R. Let. , vol 76, pp. 2511-2514, 1996

European Union
European Social Fund
OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS
MANAGING AUTHORITY
NSRF 2007-2013
EUROPEAN SOCIAL FUND
Co- financed by Greece and the European Union

On Optimization Techniques for Calibration of Stochastic Volatility Models.

Milan Mrázek, Jan Pospíšil, Tomáš Sobotka

***Abstract*—The aim of this paper is to study stochastic volatility models and their calibration to real market data. This task is formulated as the optimization problem and several optimization techniques are compared and used in order to minimize the difference between the observed market prices and the model prices. At first we demonstrate the complexity of the calibration process on the popular Heston model and we show how well the model can fit a particular set of market prices. This is ensured by using a deterministic grid which eliminates the initial guess sensitivity specific to this problem. The same level of errors can be reached by employing optimization techniques introduced in the paper, while also preserving time efficiency. We further apply the same calibration procedures to the recent fractional stochastic volatility model, which is a jump-diffusion model of market dynamics with approximative fractional volatility. The novelty of this paper is especially in showing how the proposed calibration procedures work for even more complex SV model, such as the introduced long-memory fractional model.**

***Keywords*—stochastic volatility models; Heston model; fractional SV model; option pricing; calibration; optimization**

I. Introduction

In finance, stochastic volatility (SV) models are used to evaluate derivative securities, such as options. These models were developed out of a need to modify the Nobel price winning Black Scholes model [3] for option pricing, which failed to effectively take the volatility in the price of the underlying security into account. The Black Scholes model assumed that the volatility of the underlying security was constant, while SV models consider it to be a stochastic process. Among the first publications about stochastic volatility models were Hull and White [9], Scott [19], Stein and Stein [21] and Heston [8].

Calibration is the process of identifying the set of model parameters that are most likely given by the observed data. Heston model was the first model that allowed reasonable calibration to the market option data together with semi-closed form solution for European call/put option prices. Heston model also allows correlation between the asset price and the volatility process as opposed to Stein and Stein [21]. Although the Heston model was already introduced in 1993 and several other SV models appeared, Heston model is nowadays still one of the most popular models for option pricing.

The industry standard approach in calibration is to minimize the difference between the observed prices and the model prices. Option pricing models are calibrated to prices observed on the market in order to compute prices of more complex (exotic) options or hedge ratios. The complexity of the model calibration process increases with more realistic models and the fact that the estimation method of model parameters becomes as crucial as the model itself is mentioned by Jacquier and Jarrow [11].

In our case, the input parameters can not be directly observed from the market data, thus empirical estimates are of no use. It was well documented in Bakshi et al. [2] that the model implied parameters differ significantly from their time-series estimated counterparts. Cited paper for example shows, that the magnitudes of time-series correlation coefficient of the asset return and its volatility estimated from the daily prices were much lower than their model implied counterparts.

Moreover, the information observed from market data is insufficient to exactly identify the parameters, because several sets of parameters may be performing well and provide us with model prices that are close to the prices observed on the market. This is what causes the ill-posedness of the calibration problem.

The paper is organized as follows. In section II we briefly introduce the Heston model together with the semi-closed form solution for vanilla options. In section III we demonstrate the complexity of the calibration process and employ variety of optimizers. Among the considered methods there are be two global optimizers Genetic algorithm (GA) and Simulated annealing (SA) as well the local search method (denoted by LSQ). We use the different optimizers with a specific approach to calibrate the Heston model to data obtained from the real market, namely we use daily data for DAX Options obtained using the Bloomberg Terminal and in the next section the FTSE 100 options.

In section IV we introduce a jump-diffusion model of market dynamics with approximative fractional volatility (FSV model). We show that the option pricing problem under this model attains a semi-closed form solution and demonstrate how the optimization procedures can be used for the calibration task. We will conclude our results in section V.

II. Heston model

Following Heston [8] and Rouah [18] we consider the risk-neutral stock price model

$$dS_t = rS_t dt + \sqrt{v_t} S_t d\widetilde{W}_t^S, \tag{1}$$

$$dv_t = \kappa(\theta - v_t)dt + \sigma\sqrt{v_t} d\widetilde{W}_t^v, \tag{2}$$

$$d\widetilde{W}_t^S d\widetilde{W}_t^v = \rho\, dt, \tag{3}$$

All authors are with NTIS - New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic, emails: {mrazekm,honik,sobotkat}@ntis.zcu.cz

Manuscript received November 10, 2014.

with initial conditions $S_0 \geq 0$ and $v_0 \geq 0$, where S_t is the price of the underlying asset at time t, v_t is the instantaneous variance at time t, r is the risk-free rate, θ is the long run average price variance, κ is the rate at which v_t reverts to θ and σ is the volatility of the volatility. $(\widetilde{W}^S, \widetilde{W}^v)$ is a two-dimensional Wiener process under the risk-neutral measure $\widetilde{\mathbb{P}}$ with instantaneous correlation ρ.

Stochastic process v_t is referred to as the variance process (also known as volatility process) and it is the square-root mean reverting process, CIR process Cox et al. [5]. It is strictly positive and cannot reach zero if the Feller [6] condition $2\kappa\theta > \sigma^2$ is satisfied.

Heston SV model allows for a semi-closed form solution for vanilla option, which involves numerical computation of an integral. Several pricing formulas were added to the original one by Heston [8], e.g. Albrecher et al. [1], Kahl and Jäckel [12], Lewis [14] or Zhylevskyy [25]. We will use here the formulas by Lewis [14]. Let K be the strike price and $\tau = T - t$ be the time to maturity. Then the price of a European call option at time t on a non-dividend paying stock with a spot price S_t is

$$C(S, v, t) = S - Ke^{-r\tau}\frac{1}{\pi}\int_{0+i/2}^{\infty+i/2} e^{-ikX}\frac{\hat{H}(k, v, \tau)}{k^2 - ik}dk, \tag{4}$$

where $X = \ln(S/K) + r\tau$ and

$$\hat{H}(k, v, \tau) = \exp\left(\frac{2\kappa\theta}{\sigma^2}\left[tg - \ln\left(\frac{1 - he^{-\xi t}}{1 - h}\right) + \right.\right.$$
$$\left.\left. + vg\left(\frac{1 - e^{-\xi t}}{1 - he^{-\xi t}}\right)\right]\right),$$

where

$$g = \frac{b - \xi}{2}, \quad h = \frac{b - \xi}{b + \xi}, \quad t = \frac{\sigma^2\tau}{2},$$
$$\xi = \sqrt{b^2 + \frac{4(k^2 - ik)}{\sigma^2}},$$
$$b = \frac{2}{\sigma^2}\Big(ik\rho\sigma + \kappa\Big).$$

III. Calibration of SV model

The model calibration is formulated as an optimization problem. The aim is to minimize the pricing errors between the model prices and the market prices for a set of traded options. A common approach to measure these errors is to use the squared differences between market prices and prices returned by the model, this approach leads to the nonlinear least square method

$$\inf_{\Theta} G(\Theta), \quad G(\Theta) = \sum_{i=1}^{N} w_i |C_i^{\Theta}(t, S_t, T_i, K_i) - C_i^*(T_i, K_i)|^2, \tag{5}$$

where N denotes the number of observed option prices, w_i is a weight, $C_i^*(T_i, K_i)$ is the market price of the call option observed at time t. C^{Θ} denotes the model price computed using vector of model parameters, for Heston SV model we have $\Theta = (\kappa, \theta, \sigma, v_0, \rho)$.

The function G is an objective function of the optimization problem (5) and it is neither convex nor of any particular structure. It may have more than one global minimum and it is not possible to tell whether a unique minimum can be reached by gradient based algorithm. When searching for the global minimum, a set of linear constraints must be also added to the problem, because of the parameters values. For example in Heston SV model, ρ represents correlation coefficient and thus ρ needs to only attain values within the interval $[-1, 1]$.

Local deterministic algorithms can be used to solve the calibration problem, but there is significantly high risk for them to end up in a local minimum, also initial guess needs to be provided for them, which appears to affect the performance of local optimizers severely.

Different take on the calibration is represented by the regularisation method. Penalization function, e.g., $f(\Theta)$ such that

$$\inf_{\Theta} G(\Theta) + \alpha f(\Theta)$$

is convex, is added to the objective function (5), which enables the usage of gradient based optimizing procedures. This method yields another parameter to be estimated α, which is called regularisation parameter. More details on this approach can be seen in Cont and Hamida [4].

A. Considered algorithms

Facing the calibration problem (5), we took into account several optimizing methods and tested these on a set of generated prices by the actual Heston model. That means that we were aware of the parameters that would fully explain the synthetic generated market prices and we were able to judge the algorithms on how close they were able get to this set of parameters. The ones that performed best we used afterwards on a real set of data observed on market. Among the considered methods were two (heuristic) global optimizers Genetic algorithm (GA) and Simulated annealing (SA) as well as the local search method (denoted by LSQ). GA and SA are available in MATLAB's Global Optimization Toolbox as functions `ga()` and `simulannealbnd()` respectively, whereas LSQ is available in MATLAB's Optimization Toolbox as function `lsqnonlin()` that implements the Gauss-Newton trust-region-reflective method with the possibility of choosing the Levenberg-Marquardt algorithm. We also tested performance of MS Excel's solver and Adaptive simulated annealing (ASA) (available at ingber.com) as well as modSQP suggested in Kienitz and Wetterau [13]. Based on the results we abandoned MATLAB's Simulated annealing (SA) and modSQP. They both seemed to work fine on a small example but they did not seem to be applicable for a larger number of generated strikes and maturities.

B. Measured errors

As a criterion for the performance evaluation of the optimizing methods we were recording the following errors:

ISBN: 978-1-61804-253-8

$$\text{MARE}(\Theta) = \max_i \frac{|C_i^\Theta - C_i^*|}{C_i^*} \tag{6}$$

and

$$\text{AARE}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{|C_i^\Theta - C_i^*|}{C_i^*} \tag{7}$$

for $i = 1, \ldots, N$. MARE denotes maximum absolute relative error and AARE is the average of the absolute relative error across all strikes and maturities.

C. Considered weights

Weights in (5) are denoted by w_i. It makes sense to put the most weight where the most liquid quotes are on the market, which is usually around ATM. We employed the bid ask spreads $\delta_i > 0$ with our market data and aimed to have the model prices close to the mid prices, that are considered as the market prices C_i^*. We decided not to limit ourselves with just one choice for the weight function, but to test more of these and explore any influence on the results caused by the particular choice of the weight function. The weights are denoted by capital letters A,B,C as follows:

$$\text{weight A:} \quad w_i = \frac{1}{|\delta_i|}, \tag{8}$$

$$\text{weight B:} \quad w_i = \frac{1}{\delta_i^2}, \tag{9}$$

$$\text{weight C:} \quad w_i = \frac{1}{\sqrt{\delta_i}}. \tag{10}$$

The above means that the bigger the spread the less weight is put on the particular difference between the model price and the market price (mid price) during the calibration process.

D. Empirical results for Heston model on real market data

In order to overcome the initial guess sensitivity, we chose to adopt the approach of combining the global and local optimizers. We would start with a global optimizer (GA, ASA) and provide the result as an initial guess to a local optimizer (LSQ, Excel's solver). We tested this approach on market prices obtained on March 19, 2013 from Bloomberg's Option Monitor for ODAX call options. We used a set of 107 options for 6 maturities. S_0 was the current DAX value at the time and as r we took the corresponding EURIBOR rate. For a benchmarking purpose we used an error, that was achieved by dividing the state space of all possible parameters into a grid in order to obtain a large set of initial values. Starting from all these initial values was rather time consuming, nevertheless it served us as an indicator of how well the model can actually explain this particular set of market prices. We were able to achieve AARE of 0.58%, see TABLE I.

Following the results in TABLE I, we can see that Excel failed to significantly refine the initial values provided by the global optimizers. On the other hand using LSQ we were able to refine the initial guess for the parameters provided by GA

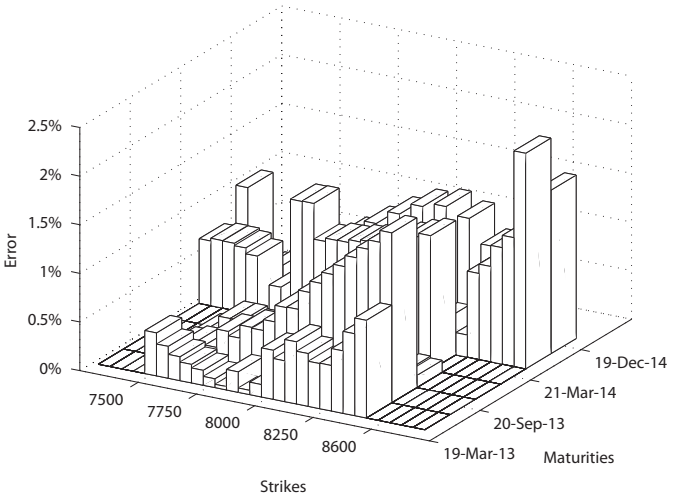


Fig. 1. Results of calibration for pair GA and LSQ in terms of absolute relative errors.

and obtain an average absolute relative error 0.65%, see Fig. 1, which is comparable to the average absolute relative error obtained by the method using deterministic grid as the initial value for LSQ mentioned above. Moreover the maximum absolute relative error was only 2.22% using the approach of combining the optimization methods.

When comparing the global optimizers GA provided better results than ASA. Also the initial guess from GA was the one which was later refined by LSQ producing lowest maximum absolute relative error of 2.22%, which can be observed in Fig. 1.

As it shows in TABLE I, the choice of weights can play a significant role during the calibration process. Different weights yielded best results for both GA and ASA, however LSQ seems to be favoring weights B. For all the results mentioned above see TABLE I and for more results see [16].

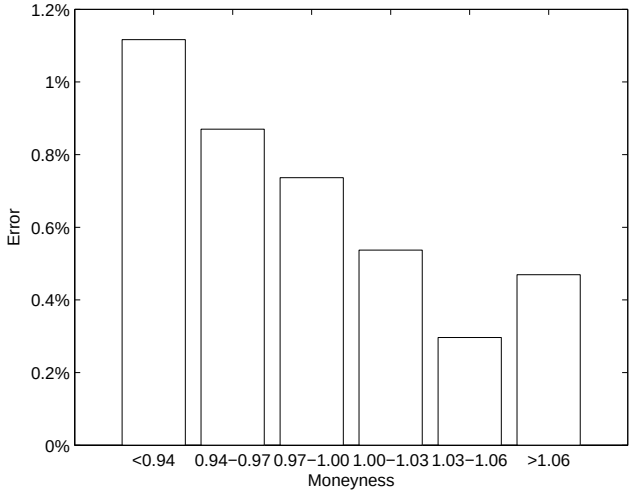


Fig. 2. Results of calibration for pair GA and LSQ in terms of average absolute relative errors.

TABLE I
CALIBRATION RESULTS FOR MARCH 19, 2013.* INITIAL GUESSES OBTAINED BY DETERMINISTIC GRID.

Algorithm	Weight	AARE	MARE	v_0	κ	θ	σ	ρ
GA	A	1.25%	12.46%	0.02897	0.68921	0.10313	0.79492	-0.53769
GA	B	2.10%	13.80%	0.03073	0.06405	0.94533	0.91248	-0.53915
GA	C	1.70%	18.35%	0.03300	0.83930	0.10826	1.14674	-0.49923
ASA	A	2.26%	19.51%	0.03876	0.80811	0.13781	1.63697	-0.46680
ASA	B	2.62%	28.65%	0.03721	1.45765	0.09663	1.86941	-0.37053
ASA	C	1.73%	19.82%	0.03550	1.22482	0.09508	1.44249	-0.49063
LSQ*	B	0.58%	3.10%	0.02382	1.75680	0.04953	0.42134	-0.84493
GA+Excel	A	1.25%	12.46%	0.02897	0.68922	0.10314	0.79490	-0.53769
GA+Excel	B	1.25%	12.46%	0.02896	0.68921	0.10314	0.79492	-0.53769
GA+Excel	C	1.25%	12.66%	0.02903	0.68932	0.10294	0.79464	-0.53763
ASA+Excel	A	1.73%	19.82%	0.03550	1.22482	0.09509	1.44248	-0.49062
ASA+Excel	B	1.78%	18.18%	0.03439	1.22399	0.09740	1.43711	-0.49115
ASA+Excel	C	1.73%	19.82%	0.03550	1.22482	0.09509	1.44248	-0.49062
GA+LSQ	A	0.67%	3.07%	0.02491	0.82270	0.07597	0.48665	-0.67099
GA+LSQ	B	0.65%	2.22%	0.02497	1.22136	0.06442	0.55993	-0.66255
GA+LSQ	C	0.68%	3.66%	0.02486	0.75195	0.07886	0.46936	-0.67266
ASA+LSQ	A	1.73%	19.82%	0.03550	1.22482	0.09508	1.44249	-0.49063
ASA+LSQ	B	1.71%	19.48%	0.03511	1.22672	0.09636	1.44194	-0.49089
ASA+LSQ	C	1.73%	19.82%	0.03550	1.22482	0.09508	1.44249	-0.49063

IV. MODEL WITH APPROXIMATIVE FRACTIONAL STOCHASTIC VOLATILITY

In this section we introduce a jump-diffusion model of market dynamics with approximative fractional volatility. We show that the option pricing problem under this model attains a semi-closed form solution and previously mentioned optimization procedures can be used for the calibration task.

A. Model introduction

We consider a model, firstly proposed by Intarasit and Sattayatham [10], that takes the following form under a risk-neutral measure

$$\begin{aligned} dS_t &= rS_t dt + \sqrt{v_t} S_t dW_t^S + Y_t S_{t-} dN_t, \quad (11)\\ dv_t &= -\kappa(v_t - \bar{v})dt + \xi v_t dB_t^H, \quad (12)\end{aligned}$$

where $\kappa, \bar{v}, \xi$ are model parameters, such that, κ is a mean-reversion rate, $\bar{v}$ stands for an average volatility level and finally, ξ is so-called volatility of volatility. Under the notation S_{t-} we understand $\lim_{s\to t-} S_s$ and $(N_t)_{t\geq 0}, (W_t^S)_{t\geq 0}$ is a Poisson process and a standard Wiener process respectively. Y_t denotes an amplitude of a jump at t (conditional on occurrence of the jump).

A stochastic process $(B_t^H)_{t\geq 0}$ can be formally defined as

$$B_t^H = \int_0^t (t-s+\varepsilon)^{H-1/2} dW_s, \quad (13)$$

where H is a long-memory parameter, ε is a non-negative approximation factor [10] and, as previously, $(W_t)_{t\geq 0}$ represents a standard Wiener process. Thao [22] showed that for $\varepsilon \to 0$, $(B_t^H)_\varepsilon$ converges uniformly to a non-Markov process and H in that case coincides with the well-known Hurst parameter ranging in $[0, 1]$. For financial applications we are interested in a long-range dependence of volatility, therefore we consider $H \in (0.5, 1]$. Moreover, if $\varepsilon > 0$ then B_t^H is a semi-martingale [24]. Hence, the Itô stochastic calculus can be used when deriving an explicit model price for European options. dB_t corresponds to the following integral which was defined for arbitrary stochastic process with bounded variation $(F_t)_{t\geq 0}$ by Nguyen & Thao [23]

$$\int_0^t F_s dB_s := F_t B_t - \int_0^t B_s dF_s - [F, B]_t, \quad (14)$$

provided the right-hand side integral exists in a Riemann-Stieltjes sense, while $[F, B]_t$ being a quadratic variation of $F_t B_t$. Under this setup we are able to rewrite the original system of stochastic differential equations into

$$\begin{aligned} dS_t &= rS_t dt + \sqrt{v_t} S_t dW_t^S + Y_t S_{t-} dN_t, \quad (15)\\ dv_t &= \alpha dt + \beta dW_t^v, \quad (16)\end{aligned}$$

with the drift process $\alpha := \alpha(S_t, v_{t,t}) = (a\xi\varphi_t - \kappa)v_t + \theta$ and diffusion $\beta := \beta(S_t, v_{t,t}) = \xi v_t \varepsilon^a$, where $a := H - 1/2$, $\theta := \kappa\bar{v}$ is a constant and φ_t represents an Itô integral,

$$\varphi_t = \int_0^t (t-s+\varepsilon)^{H-3/2} dW_s^\varphi \quad (17)$$

alongside standard Wiener processes $(W_t^v)_{t\geq 0}, (W_t^\varphi)_{t\geq 0}$. We add an instantaneous correlation $\mathbb{E}[dW_t^S dW_t^v] = \rho dt$ to mimic the stock-volatility leverage effect and we also consider a jump process with log-normally distributed jump sizes. Jump times are due to Poisson process $(N_t)_{t\geq 0}$ with parameter λ. To simplify the pricing problem, $Y_t dN_t$ is set to be stochastically independent on processes driving SDE's (15)-(16).

The above described setting is referred to as the FSV model throughout this text. In the calibration problem (5) for the FSV model, the vector of parameters to be optimized will be $\Theta =$

$(v_0, \kappa, \bar{v}, \xi, \rho, \lambda, \alpha_J, \gamma_J, H)$. Their meaning is summarized in TABLE II.

TABLE II
LIST OF FSV MODEL PARAMETERS.

v_0 initial volatility	κ mean reversion rate	$\bar{v}$ average volatility
ξ volatility of volatility	ρ correlation coef.	λ Poisson hazard rate
α_J expected jump size	γ_J variance of jump sizes	H Hurst parameter

B. Semi-closed form solution

In what follows, we present a semi-closed form solution to the pricing problem for European options. The solution is about to be derived for pure-diffusion dynamics first, then the jump part is considered.

We focus on a European call option expiring at time T with pay-off $(S_T - K)^+$ where K is a strike price of the contract. The modeled price V should equal to a discounted expected pay-off under a risk-neutral measure,

$$\begin{aligned} V(\tau, K) &= e^{-r\tau}\mathbb{E}\left[(S_T - K)^+\right] \\ &= S_t P_1(S_t, v_t, \tau) - e^{-r\tau} K P_2(S_t, v_t, \tau) \\ &= e^{x_t} P_1(x_t, v_t, \tau) - e^{-r\tau} K P_2(x_t, v_t, \tau). \end{aligned} \quad (18)$$

In (18) we expressed V using a time to maturity $\tau := T - t$ and logarithm of the underlying price $x_t := \ln(S_t)$. $P_1,\ P_2$ can be interpreted as the risk-neutral probabilities that option expires in the money conditional on the value of x_t and finally r is assumed to be a uniquely determined risk-free rate constant. We will retrieve $P_1,\ P_2$ in terms of characteristic functions $f_n = f_n(\phi, \tau), n = 1, 2$

$$P_n = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \Re\left[\frac{e^{i\phi\ln(K)} f_n}{i\phi}\right] d\phi, \quad (19)$$

where $\Re(z)$ denotes a real part of the complex number z and i denotes the imaginary unit. Following the original article by Heston [8], characteristic functions are to be found as

$$f_n = \exp\left\{C_n(\tau, \phi) + D_n(\tau, \phi) v_0 + i\phi x\right\}. \quad (20)$$

The pricing problem can be formulated using partial differential equations as the initial value problem [20]

$$\begin{aligned} -\frac{\partial V}{\partial \tau} + \frac{1}{2} v_t \frac{\partial^2 V}{\partial x_t^2} + \left(r - \frac{1}{2} v_t\right)\frac{\partial V}{\partial x_t} + \rho\beta v_t \frac{\partial^2 V}{\partial v_t \partial x_t} \\ - rV + \frac{1}{2} v_t \beta^2 \frac{\partial^2 V}{\partial v_t^2} + \alpha \frac{\partial V}{\partial v_t} = 0; \quad (21) \end{aligned}$$

$$V(0, K) = (S_T - K)^+. \quad (22)$$

We are able to split (21) into two equations with respect to $P_1,\ P_2$:

$$\begin{aligned} -\frac{\partial P_1}{\partial \tau} + \frac{1}{2} v_t \frac{\partial^2 P_1}{\partial x_t^2} + \left(r + \frac{1}{2} v_t\right)\frac{\partial P_1}{\partial x_t} + \rho\beta v_t \frac{\partial^2 P_1}{\partial v_t \partial x_t} \\ + \frac{1}{2} v_t \beta^2 \frac{\partial^2 P_1}{\partial v_t^2} + (\alpha + \rho\beta v_t)\frac{\partial P_1}{\partial v_t} = 0. \end{aligned} \quad (23)$$

$$\begin{aligned} -\frac{\partial P_2}{\partial \tau} + \frac{1}{2} v_t \frac{\partial^2 P_2}{\partial x_t^2} + \left(r - \frac{1}{2} v_t\right)\frac{\partial P_2}{\partial x_t} + \rho\beta v_t \frac{\partial^2 P_2}{\partial v_t \partial x_t} \\ + \frac{1}{2} v_t \beta^2 \frac{\partial^2 P_2}{\partial v_t^2} + \alpha \frac{\partial P_2}{\partial v_t} = 0. \end{aligned} \quad (24)$$

Using arguments in [17], [20] and a discounted version of the Feynman-Kac theorem, characteristic functions (20) satisfy the previous equations and thus we can substitute $f_n,\ n = 1, 2$ into (23), (24) respectively. Combining $\alpha = (a\xi\varphi_t - \kappa)v_t + \theta$ alongside equations expressed in terms of (20), we arrive at the modified initial value problem

$$\begin{aligned} \frac{\partial D_1}{\partial \tau} &= \rho\beta i\phi D_1 - \frac{1}{2}\phi^2 + \frac{1}{2}\beta^2 D_1^2 + \frac{1}{2} i\phi \\ &\quad + (a\xi\varphi_t - \kappa + \rho\beta) D_1; \quad (25) \end{aligned}$$

$$\frac{\partial D_2}{\partial \tau} = \rho\beta i\phi D_2 - \frac{1}{2}\phi^2 + \frac{1}{2}\beta^2 D_2^2 - \frac{1}{2} i\phi + (a\xi\varphi_t - \kappa) D_2; \quad (26)$$

$$\frac{\partial C_n}{\partial \tau} = ri\phi + \theta D_n; \quad (27)$$

with the initial condition

$$C_n(0, \phi) = D_n(0, \phi) = 0. \quad (28)$$

The first two equations for D_n are known as the Riccati equations with constant coefficients. Once D_n are obtained, the last two ordinary differential equations are solved by a direct integration.

Proposition 1. *The characteristic functions of the logarithmic stock price $f_n = f_n(\tau, \phi)$ under the FSV model take the form*

$$f_n = \exp\left\{C_n(\tau, \phi) + D_n(\tau, \phi) v_0 + i\phi \ln(S_t) + \psi(\phi)\tau\right\}$$

with

$$\begin{aligned} C_n(\tau, \phi) &= r\phi i\tau + \theta Y_n \tau - \frac{2\theta}{\beta^2}\ln\left(\frac{1 - g_n e^{d_n\tau}}{1 - g_n}\right), \\ D_n(\tau, \phi) &= Y_n\left(\frac{1 - e^{d_n\tau}}{1 - g_n e^{d_n\tau}}\right), \\ \psi &= -\lambda i\phi\left(e^{\alpha_J + \gamma_J^2/2} - 1\right) + \lambda\left(e^{i\phi\alpha_J - \phi^2\gamma_J^2/2} - 1\right) \\ Y_n &= \frac{b_n - \rho\beta\phi i + d_n}{\beta^2} \\ g_n &= \frac{b_n - \rho\beta\phi i + d_n}{b_n - \rho\beta\phi i - d_n}, \\ d_n &= \sqrt{(\rho\beta\phi i - b_n)^2 - \beta^2(2u_n\phi i - \phi^2)}, \\ \beta &= \xi\varepsilon^{H - 1/2}\sqrt{v_t},\ u_1 = 1/2,\ \ u_2 = -1/2,\ \theta = \kappa\bar{v}, \\ b_1 &= \kappa - (H - 1/2)\xi\varphi_t - \rho\beta, \\ b_2 &= \kappa - (H - 1/2)\xi\varphi_t. \end{aligned}$$

Outline of the proof. Firstly, we solve the initial value problem (25)-(27), which gives us characteristic functions under the FSV model without jumps (see [20]). Using stochastic independence of the jump process and diffusion processes, the joint characteristic function is obtained as a product of

individual characteristic functions [7]. Particular steps of the proof can be found in [17]. □

Characteristic functions f_n are used in expression (19) to evaluate the risk-neutral probabilities P_n. The integral was computed, in our case, by an adaptive Lobatto quadrature, implemented in MATLAB's `quadl()` function. When P_n corresponding to the given European call are obtained, the price is retrieved by expression (18).

C. Calibration results

The formula is an explicit version of the one in [10] and, as we illustrate on market data, it is ready to be used out of the box. To compare the new FSV approach with the Heston model introduced in previous sections, we utilized market data on the British FTSE 100 stock index (8^{th} January 2014). The index was quoted at 6,721.80 points and our main data set consisted of 82 traded options. In terms of moneyness, considered options are of both in-the-money, at-the-money and out-of-the-money.

TABLE III
CALIBRATION ERRORS. FTSE 100 OPTION MARKET, **WEIGHTS A**, DATA SET OBTAINED ON 8^{th} JANUARY 2014.

Model	Algorithm	AARE [%]	MARE [%]
FSV model	GA+LSQ	2.34	20.53
	SA+LSQ	2.34	20.53
Heston model	GA+LSQ	3.36	19.01
	SA+LSQ	4.43	29.34

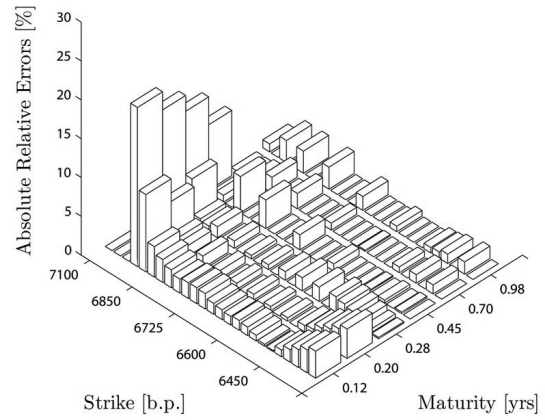
For the model comparison, global optimizers (GA, SA) were used to obtain initial guess for a trust-region local search method (LSQ). The FSV model provided slightly better average market errors and also was more consistent throughout different sets of weights. However, the optimization problem (5) is more complicated for this approach,which is caused by having four more parameters to calibrate compared to the Heston model.

TABLE IV
CALIBRATION ERRORS. FTSE 100 OPTION MARKET, **WEIGHTS B**, DATA SET OBTAINED ON 8^{th} JANUARY 2014.

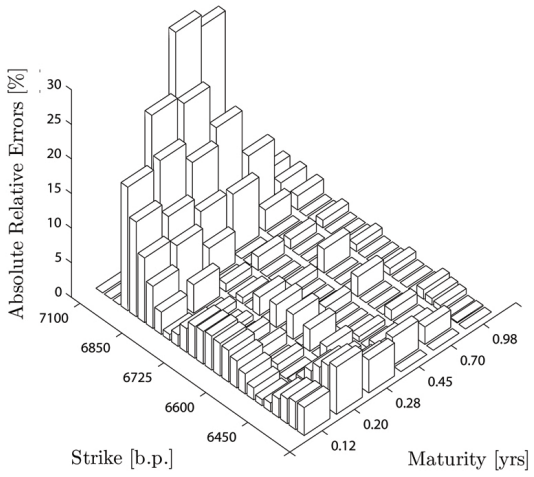
Model	Algorithm	AARE [%]	MARE [%]
FSV model	GA+LSQ	2.33	20.49
	SA+LSQ	2.34	20.53
Heston model	GA+LSQ	5.07	32.36
	SA+LSQ	4.15	23.33

TABLE V
CALIBRATION ERRORS. FTSE 100 OPTION MARKET, **WEIGHTS C**, DATA SET OBTAINED ON 8^{th} JANUARY 2014.

Model	Algorithm	AARE [%]	MARE [%]
FSV model	GA+LSQ	2.34	20.53
	SA+LSQ	2.34	20.53
Heston model	GA+LSQ	3.35	18.85
	SA+LSQ	3.52	19.93



(a) FSV model



(b) Heston model

Fig. 3. Calibration from FTSE 100 call option market using Genetic Algorithm combined with a local search method. Displayed average relative errors were obtained for weights B.

V. CONCLUSION

In this paper, we compared several optimization approaches on the problem of option market calibration. Firstly, we summoned a very popular model of market dynamics - the Heston model. The corresponding optimization problem is non-convex and may contain many local minima, hence any local search method without a good initial guess may fail to achieve satisfactory results. To overcome this well-known calibration issue, we set a fine deterministic grid for initial starting points. The best result of a trust region minimizer for these points (AARE =0.58%) is taken as a reference point for comparison of less heuristic and more efficient approaches.

Using Genetic Algorithm (GA) combined with a local optimizer (LSQ) we were able to get close to the reference result

(AARE =0.65%) in terms of the average absolute relative error and even achieve superior maximum absolute error. All used methods were also tested for different sets of weights, all of which should emphasize more liquid market contracts by using bid-ask spread.

Last but not least, we took a closer look at the calibration problem with respect to the newly proposed approximative fractional volatility model. Beforehand, we showed that this model attained a 'Heston-like' semi-closed formula. This means only a single integral expression needs to be numerically evaluated in order to obtain a European option price. Up to now, the fractional model was purely theoretically justified, whereas we provided an empirical calibration results applying the proposed optimization technique satisfactorily and also presented a comparison with the Heston model on the FTSE 100 index option market.

Investigation of optimization techniques for calibration of stochastic volatility models is an ongoing research. Possible performance and accuracy improvements of Gauss-Newton methods involve for example precalculation of gradients or Hessian matrix which is rather complicated task even in Heston model. Another possibility is to use the variable metric methods for nonlinear least squares as they are introduced in [15]. Complexity of the FSV model then opens space for fine tuning the global optimizers whose implementation in parallel and distributed computing environments is a further issue.

ACKNOWLEDGMENT

This work was supported by the GACR Grant 14-11559S Analysis of Fractional Stochastic Volatility Models and their Grid Implementation. The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) and access to the CERIT-SC computing and storage facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations (CZ.1.05/3.2.00/08.0144), is greatly appreciated.

REFERENCES

[1] H. Albrecher, P. Mayer, W. Schoutens, and J. Tistaert, "The little Heston trap," *Wilmott Magazine*, no. Jan/Feb, pp. 83–92, 2007.

[2] G. Bakshi, C. Cao, and Z. Chen, "Empirical performance of alternative option pricing models," *The Journal of Finance*, vol. 52, no. 5, pp. 2003–2049, 1997.

[3] F. Black and M. S. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, no. 3, pp. 637–54, May-June 1973.

[4] R. Cont and S. B. Hamida, "Recovering volatility from option prices by evolutionary optimization," *Journal of Computional Finance*, vol. 8, no. 4, p. 4376, 2005.

[5] J. C. Cox, J. E. Ingersoll, and S. A. Ross, "A theory of the term structure of interest rates," *Econometrica*, vol. 53, no. 2, pp. 385–407, 1985.

[6] W. Feller, "Two singular diffusion problems," *The Annals of Mathematics*, vol. 54, no. 1, pp. 173–182, 1951.

[7] J. Gatheral, *The Volatility Surface: A Practitioner's Guide*, ser. Wiley Finance. John Wiley & Sons, 2006.

[8] S. L. Heston, "A closed-form solution for options with stochastic volatility with applications to bond and currency options," *Review of Financial Studies*, vol. 6, no. 2, pp. 327–343, 1993.

[9] J. C. Hull and A. D. White, "The pricing of options on assets with stochastic volatilities," *Journal of Finance*, vol. 42, no. 2, pp. 281–300, June 1987.

[10] A. Intarasit and P. Sattayatham, "An approximate formula of European option for fractional stochastic volatility jump-diffusion model," *Journal of Mathematics and Statistics*, vol. 7, no. 3, pp. 230–238, 2011.

[11] E. Jacquier and R. Jarrow, "Bayesian analysis of contingent claim model error," *Journal of Econometrics*, vol. 94, no. 12, pp. 145 – 180, 2000.

[12] C. Kahl and P. Jäckel, "Not-so-complex logarithms in the Heston model," *Wilmott Magazine*, p. 94103, 2005.

[13] J. Kienitz and D. Wetterau, *Financial Modelling: Theory, Implementation and Practice with MATLAB Source*, ser. The Wiley Finance Series. Wiley, 2012.

[14] A. L. Lewis, *Option valuation under stochastic volatility, with Mathematica code*. Finance Press, Newport Beach, CA, 2000.

[15] L. Lukšan and E. Spedicato, "Variable metric methods for unconstrained optimization and nonlinear least squares," *J. Comput. Appl. Math.*, vol. 124, no. 1-2, pp. 61–95, 2000, numerical analysis 2000, Vol. IV, Optimization and nonlinear equations.

[16] M. Mrázek and J. Pospíšil, "Calibration and simulation of Heston model," *Applied Stochastic Models in Business and Industry*, 2014, submitted.

[17] J. Pospíšil and T. Sobotka, "Market calibration under a long memory stochastic volatility model," *Applied Mathematical Finance*, 2014, submitted.

[18] F. D. Rouah, *The Heston Model and its Extensions in Matlab and C#, + Website*, ser. Wiley Finance Series. Wiley, 2013.

[19] L. O. Scott, "Option pricing when the variance changes randomly: Theory, estimation, and an application," *The Journal of Financial and Quantitative Analysis*, vol. 22, no. 4, pp. 419–438, 1987.

[20] T. Sobotka, "Stochastic and fractional stochastic volatility models," Master's thesis, University of West Bohemia, 2014.

[21] J. Stein and S. E., "Stock price distributions with stochastic volatility: An analytic approach," *Review of Financial Studies*, vol. 4, no. 4, pp. 727–752, 1991.

[22] T. H. Thao, "An approximate approach to fractional analysis for finance," *Nonlinear Analysis: Real world Applications*, vol. 7, pp. 124–132, 2006.

[23] T. H. Thao and T. T. Nguyen, "Fractal Langevin equation," *Vietnam Journal Mathematics*, vol. 30, no. 1, pp. 89–96, 2002.

[24] M. Zähle, "Integration with respect to fractal functions and stochastic calculus. I," *Probab. Theory Related Fields*, vol. 111, no. 3, pp. 333–374, 1998.

[25] O. Zhylyevskyy, "Efficient pricing of European-style options under Heston's stochastic volatility model," *Theoretical Economics Letters*, vol. 2, no. 1, pp. 16–20, 2012.

Numerical simulation of flow over a helicopter rotor blade airfoil with a filled cavity

Constantin Rotaru, Ionică Cîrciu, and Mihai Ivănică

***Abstract*—**This paper examines the aerodynamic performances of a helicopter rotor blade which has a filled cavity on the upper surface, in order to generate more circulation. Taking into account that the advancing blade operates at low angle of attack but at high subsonic or transonic conditions, whereas the retreating blade operates at low Mach numbers and high lift coefficients, this new airfoil type could improve the lifting capability of the rotor blade and may lead to new rotors optimized for greater performances in both hover and high speed forward flight. The proposed solution consists in an airfoil with filled cavity, where the filled body is a rotating cylinder. The effect on the flow around the airfoil is the generation of vortices that reduce the flow separation downstream of the cavity. The CFD results were compared with those obtained by panel method and this suggests the possibility of a delay in flow separation on the upper surface for the retreating blade.

***Keywords*—**airfoil, helicopter aerodynamic, panel method, rotor blade, vortex strength.

I. Introduction

ONE of the most important characteristics used to judge the performance of an airfoil is the maximum static lift capability. This is a quantity that is not easily predicted even with computational methods and experimental measurements. Even from an experimental perspective, absolute values of $C_{l\,max}$ are difficult to guarantee with high precision and especially between tests performed in different wind tunnels. The maximum lift that can be developed by an airfoil when operating at a steady angle of attack is related to the type of stall characteristic of that airfoil. At low speeds, airfoils generally fall into three static stall categories, namely thin airfoil stall, leading edge stall and trailing edge stall. The measurements show that thin airfoil and leading edge stalls can be fairly sensitive to changes in airfoil shape, whereas trailing edge stall is insensitive. Most conventional helicopter rotor airfoils fall into the category of trailing edge or leading edge stall types at low to moderate Mach numbers. It is also common for a mixed stall behavior to occur on some airfoils which is a stall characteristic that is not clearly one type or another [1].

Airfoils designed for helicopter applications have traditionally been obtained through a long evolutionary process in which various levels of theory and experimental measurements have been combined in the pursuit of airfoil shapes with higher values of maximum lift, better lift-to-drag ratios, lower pitching moments and higher drag divergence Mach numbers. In general, these requirements are conflicting, making the design of general purpose rotor airfoils extremely challenging. Instead, various families of airfoils have been developed and optimized to meet the specific needs of different parts of the rotor blade. The use of different airfoils along the blade is made easier because of computer-aided design and composite manufacturing capability which involves only small additional costs over blade made with a single airfoil section.

The selection of airfoil sections for helicopter rotors is more difficult than for a fixed-wing aircraft because they are not point designs. For angle of attack and Mach number vary continuously at all blade elements on the rotor and one airfoil section cannot meet all the various aerodynamic requirements.

The rotor limits may be determined by either advancing blade compressibility effects or retreating blade stall. Because the onset of flow separation may limit rotor performance, there has been a great deal of emphasis in rotor design on maximizing the lifting capability of rotor airfoil sections to simultaneously alleviate both compressibility effects and retreating blade stall. The rotor design point must recognize the influence of both effects as limiting factors as well as allow sufficient margins from the stall/compressibility boundary for perturbations in angle of attack and Mach number associated with maneuvering flight and turbulent air.

The aerodynamic characteristics of rotor airfoils must be assessed at their actual operational Reynolds numbers and Mach numbers. The maximum lift coefficient, $C_{l\,max}$, can be used as one indicator of the significance of viscous effects. At the low end of the practical Reynolds number range for rotors, most airfoils have relatively low values of $C_{l\,max}$. This is because the viscous forces are more determinant, the boundary layer is thicker and the flow will separate from the airfoil surface [2].

At higher angles of attack the adverse pressure gradients produced on the upper surface of the airfoil result in a progressive increase in the thickness of the boundary layer and cause some deviation from the linear lift versus angle of attack

Constantin Rotaru is with the Aviation Integrated Systems and Mechanics Department, Military Technical Academy, Bucharest, 050141, Romania (corresponding author, tel: +40745974488; fax: +4021 335 57 63; e-mail: rotaruconstantin@yahoo.com).

Ionică Cîrciu is with the Department of Aviation, "Henri Coandă" Air Force Academy, Braşov 500187, Romania (e-mail: circiuionica@yahoo.co.uk).

Mihai Ivănică is with the Aviation Integrated Systems and Mechanics Department, Military Technical Academy, Bucharest, Romania (e-mail: ivanica.mihai@gmail.com).

behavior. On many airfoils, the onset of flow separation and stall occurs gradually with increasing of angle of attack but on some airfoil (those with sharp leading edges), the flow separation may occur quite suddenly. In the stalled flow regime, the flow over the upper surface of the airfoil is characterized by a region of fairly constant static pressure. The pitching moment about ¼-chord is much more negative because with the almost constant pressure over the upper surface the center of pressure is close to mid-chord. Less lift is generated by the airfoil because of the reduction in circulation and loss of suction near the leading edge and the drag is greater. Under these separated flow conditions, steady flow no longer prevails, with turbulence and vortices being ahead alternately from the leading and trailing edges of the airfoil into the wake [3].

The envelope of rotor thrust limits is the outcome of operation on the blades of stall effects at high angle of incidence and compressibility effects at high Mach number. Usually the restrictions occur within the limits of available power. In hover, conditions are uniform around the azimuth and blade stall sets a limit to the thrust available. As forward speed increases, maximum thrust on the retreating blade falls because of the drop in dynamic pressure and this limits the thrust achievable throughout the forward speed range. By the converse effect, maximum thrust possible on the advancing side increases but is unrealizable because of the retreating blade restriction. At higher speeds, as the advancing tip Mach number approaches 1.0, its lift becomes restricted by shock-induced flow separation, leading to drag or pitching moment divergence, which limits the maximum speed achievable. Thus, the envelope comprises a limit on thrust from retreating blade stall and a limit on forward speed from advancing blade Mach effects [4].

The ability to develop computers methods in performance calculation has been a major factor in the rapid development of helicopter technology. Results may often not be greatly different from those derived from the simple analytical formulae but the fact that the feasibility of calculation is not dependent upon making a large number of challengeable assumptions is important in pinning down a design, making comparisons with flight tests [5].

II. Helicopter Rotor Blade Aerodynamic

The rotor limits may be determined by two conditions, one codition given by advancing blade compressibility effects and the other one condition given by retreating blade stall. In either case the advancing blade operates at low angle of attack but at high subsonic or transonic conditions, whereas the retreating blade operates at low Mach numbers and high lift coefficients [6].

The region of the rotor disk affected by compressibility effects is shown in fig. 1 and is defined on the surface where the incident Mach number of the flow that is normal to the leading edge of the blade exceeds the drag divergence Mach number, M_{dd}. If $M_{\Omega R}$ is the hover tip Mach number, than the region of the disk affected by compressibility effects is defined by

$$M_{r,\psi} = M_{\Omega R}(r + \mu \sin\psi) \geq M_{dd} \quad (1)$$

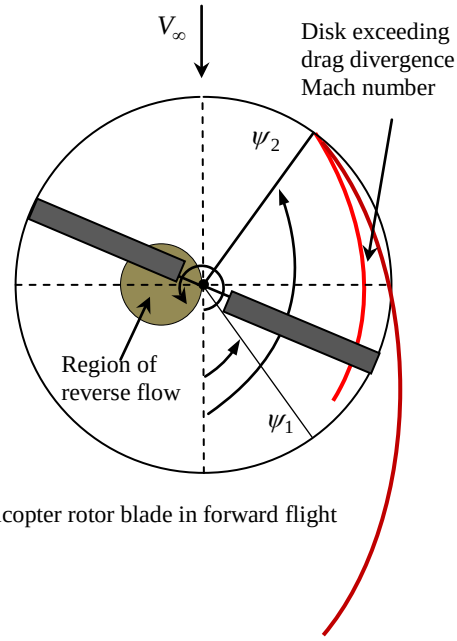


Fig. 1 helicopter rotor blade in forward flight

The angular or rotational speed of the rotor is denoted by Ω, the rotor radius by R, the advanced ratio $\mu = V_\infty \cos\alpha / \Omega R$ and $r = y / R$ where y is the axis along the rotor blade and α is the angle between the forward velocity V_∞ and the plane of the rotor [1, 2]. The azimuth angle for the onset drag divergence, ψ_1, can be obtained by setting $r = 1$, so that

$$\psi_1 = arcsin\left(\left[\frac{1}{\mu}\left(\frac{M_{dd}}{M_{\Omega R}} - 1\right)\right]\right) \quad (2)$$

and $\psi_2 = 180 - \psi_1$.

Another complication with helicopter rotors is that the wakes and tip vortices from other blades can lie close to each other and to the plane of blade rotation and so they have large induced effects on the blade lift distribution (fig. 2).

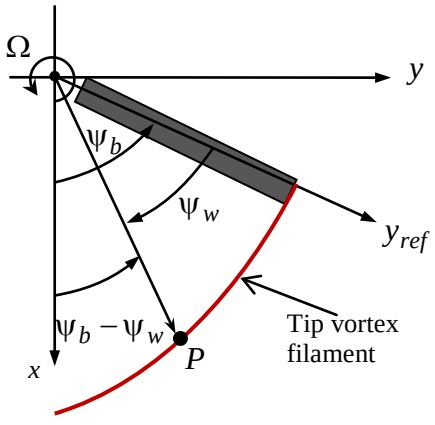


Fig. 2 tip vortex trajectory

If the wake is assumed to be undistorted in the tip path plane and no wake contraction occurs in the radial direction, then the tip vortex trajectories are described by the equations

$$\begin{cases} x = R\cos(\psi_b - \psi_w) + R\mu\psi_w \\ y = R\sin(\psi_b - \psi_w) \end{cases} \quad (3)$$

where ψ_b is the position of the blade when the vortex was formed and ψ_w is the position of the vortex element relative to the blade. These interactions of blades and tip vortices (called blade-vortex-interactions) can occur at many different locations over the rotor disk and also with different orientations.

III. Airfoil Blade with Filled Cavity

Two-dimensional simulations were performed for a standard NACA 2412 airfoil with and without cavity. Both edges of the cavity are sharp in order to fix the separation point (forward edge) and to maximize the feedback loop of the shear layer (rear edge). The cavity was filled with a rotating small cylinder for improving the circulation around the airfoil (fig. 3).

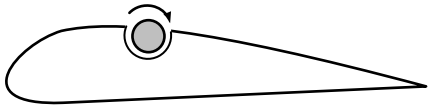


Fig. 3 airfoil with filled cavity

The computational domain extends to a distance of 12 chords lengths in the upstream and downstream directions and three chords lengths in the upper and lower normal directions. The distance between the discrete points at which the non-slip condition is enforced needs to be equal to or slightly greater than the grid spacing. The grid resolution and domain size were varied in order to assess convergence and influence of the far-field boundary condition. The Reynolds number was sufficiently high such that the formation of large scale vortices and the subsequent pairing of these structures gives rise to aperiodic low frequency oscillations that are difficult to characterize because the run times are not sufficiently long to observe many periods.

The CFD results are presented in the figures 4-7.

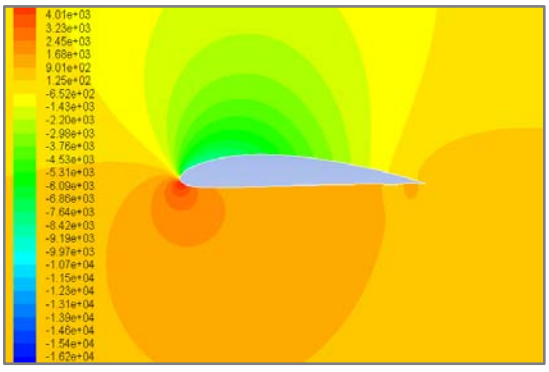


Fig. 4 airfoil without cavity: pressure distribution

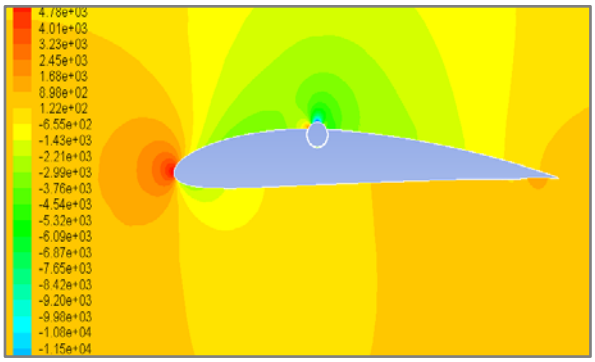


Fig. 5 airfoil with filled cavity: pressure distribution

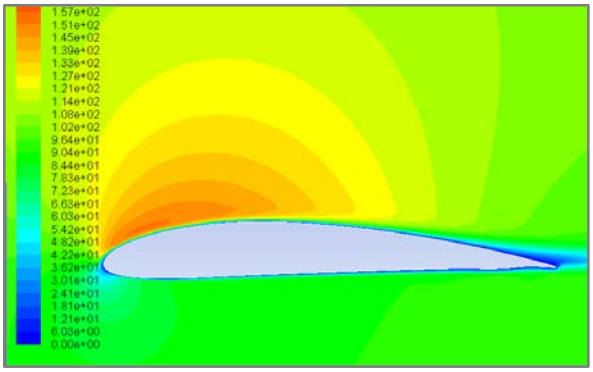


Fig. 6 airfoil without cavity: velocity distribution

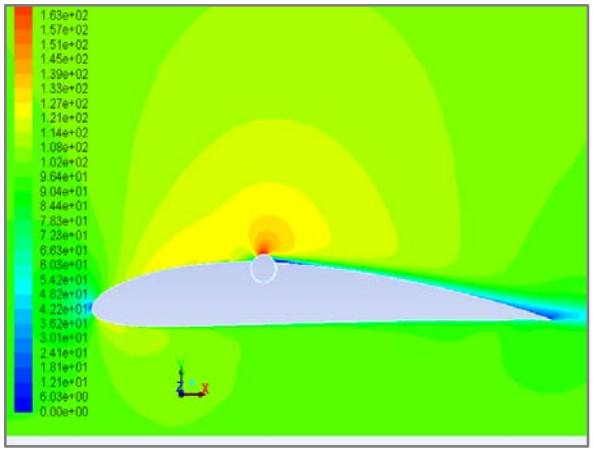


Fig. 7 airfoil with filled cavity: velocity distribution

The relative high thickness of the airfoil without a cavity causes a laminar separation which initially starts approximately half a chord length from the leading edge. At very high angles of attack the flow over the airfoil with cavity separates well before the forward edge of the cavity. The separated flow displays a strong interaction with the cavity and this interaction causes the flow to shed smaller scale structures than the airfoil without cavity at the same angle of attack.

ISBN: 978-1-61804-253-8

IV. Panel Method Results

Potential flow over an airfoil of arbitrary shape can be synthesized by combining uniform flow with a curved vortex sheet wrapped around the surface of the airfoil. The concept of replacing the airfoil surface with a vortex sheet is more than just a mathematical device because there is a thin boundary layer on the surface, due to the action of friction between the surface and the airflow, in which the large velocity gradients produce substantial vorticity, hence, there is a distribution of vorticity along the airfoil surface due to viscous effects [7].

The vortex strength, $\gamma(s)$ must vary along the surface such that the normal component of velocity induced by the entire sheet and the uniform flow is zero everywhere along the surface of the airfoil. In most cases, the strength distribution necessary to satisfy this condition is difficult to be determined analytically. For numerical computations, such sheet can be approximated as a series of flat vortex panels wrapped around the surface of the airfoil (fig. 8).

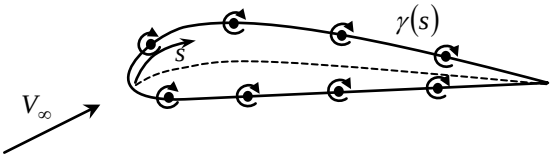


Fig. 8 vortex sheet

To define the vortex panels, a series of nodes is placed on the airfoil surface, such that the nodes are clustered more tightly near the leading and trailing edges. The change of variable $x/c=(1-cos\theta)/2$ provides the desired clustering in *x*.

The panels start at the trailing edge, are spaced forward along the lower surface, are wrapped up around the leading edge and then run back along the upper surface to the trailing edge so that the last panel ends at the trailing edge where the first panel began [8]. The vortex strength $\gamma(s)$ of each panel is assumed to be linear along the panel and continuous from one panel to the next. That is, for the *n* panels, the vortex panel strengths are $\gamma_1,\gamma_2,.....\gamma_n$, and the main thrust of the panel technique is to solve for γ_j, $j=1$ to *n*, such that the body surface becomes a streamline of the flow and such that the Kutta condition $\gamma_1=-\gamma_n$ is satisfied (fig. 9).

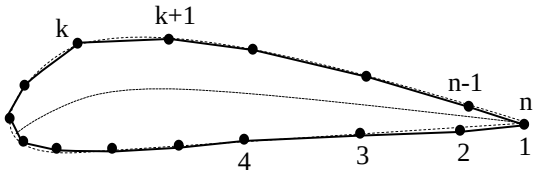


Fig. 9 vortex panel distribution

To solve for the *n* unknown nodal vortex strengths, at the center of each panel is defined a control point where the normal component of the flow velocity is imposed to be zero.

For an even number *n* of nodes, the points x_i, $i=1,2,.....n/2$ on the chord line are computed from the following algorithm:

$$\delta\theta=\frac{2\pi}{n-1},\ x_i=\frac{c}{2}\left\{1-cos\left[\left(i-\frac{1}{2}\right)\delta\theta\right]\right\},\ i=1,2,3,.....n/2$$

The lover and upper surface coordinates for an airfoil can be obtained from the camber line geometry, $y_c(x)$, and the thickness distribution, $t(x)$ as follows

$$\begin{cases} X_l(x_i)=x_i+\dfrac{t(x_i)}{2\sqrt{1+\left(\dfrac{dy_c(x)}{dx}\right)^2\Big|_{x=x_i}}}\dfrac{dy_c(x)}{dx}\Big|_{x=x_i} \\ Y_l(x_i)=y_c(x_i)-\dfrac{t(x_i)}{2\sqrt{1+\left(\dfrac{dy_c(x)}{dx}\right)^2\Big|_{x=x_i}}} \end{cases} \quad (4)$$

$$\begin{cases} X_u(x_i)=x_i-\dfrac{t(x_i)}{2\sqrt{1+\left(\dfrac{dy_c(x)}{dx}\right)^2\Big|_{x=x_i}}}\dfrac{dy_c(x)}{dx}\Big|_{x=x_i} \\ Y_u(x_i)=y_c(x_i)+\dfrac{t(x_i)}{2\sqrt{1+\left(\dfrac{dy_c(x)}{dx}\right)^2\Big|_{x=x_i}}} \end{cases} \quad (5)$$

For a point x_i on the chord line (fig. 10) we have two nodes on the airfoil, one node on the lower line of the airfoil, $P_{\frac{n}{2}+1-i}[X_l(x_1),Y_l(x_i)]$ and the other one on the upper line of the airfoil, $P_{\frac{n}{2}+i}[X_u(x_1),Y_u(x_i)]$.

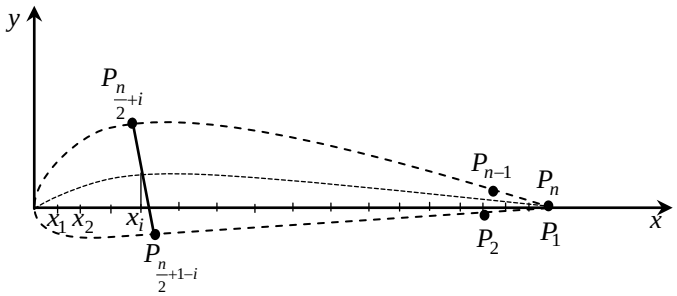


Fig. 10 the upper and lower lines nodes

A second-order panel method assumes a linear variation of $\gamma(s)$ over a given panel and the value of $\gamma(s)$ at the edges of each panel is matched to its neighbors (fig.11). The flow-tangency boundary condition is still applied at the control

point to each panel.

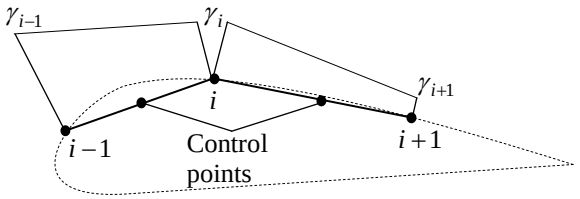


Fig. 11 linear distribution of $\gamma(s)$

The coordinates of these control points are given by

$$\begin{cases} X_C(i) = \dfrac{X_{P_i} + X_{P_{i+1}}}{2} \\ Y_C(i) = \dfrac{Y_{P_i} + Y_{P_{i+1}}}{2} \end{cases} \tag{6}$$

Each panel is assigned a local panel coordinate system (ξ, η) as shown in fig. 12.

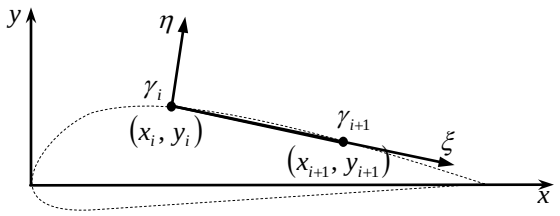


Fig. 12 vortex panel coordinate system

For each panel, an infinite number of infinitesimally weak vortices are combined in side-by-side fashion as shown in fig. 13.

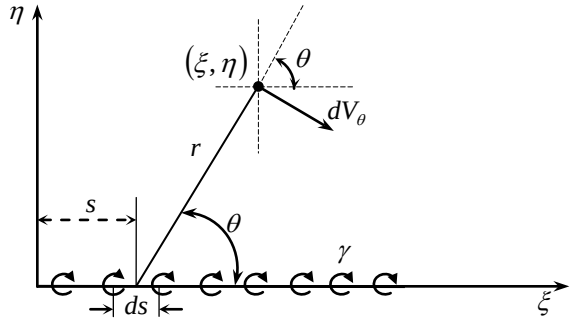


Fig. 13 edge view of a 2-D vortex panel

Consider a differential segment of a vortex panel that lies on the ξ axis at the location $\xi = s$ and has length ds. The velocity induced at any point (ξ, η) by this differential vortex is normal to the vector $\vec{r}$ and has a magnitude inversely proportional to the distance between the points of coordinates $(s, 0)$ and (ξ, η), namely $r = |\vec{r}|$. The ξ - and η - components of the velocity induced at the point (ξ, η) by this infinitesimally vortex panel are given by

$$\begin{cases} dV_\xi = dV_\theta \sin\theta = \dfrac{\gamma(s)}{2\pi r} \sin\theta \, ds \\ dV_\eta = -dV_\theta \cos\theta = -\dfrac{\gamma(s)}{2\pi r} \cos\theta \, ds \end{cases} \tag{7}$$

According to fig. 13 we have

$$\begin{cases} \sin\theta = \dfrac{\eta}{r} \\ \cos\theta = \dfrac{\xi - s}{r} \end{cases} \tag{8}$$

where $r = \sqrt{(\xi - s)^2 + \eta^2}$.

It follows that

$$\begin{cases} dV_\xi = \dfrac{\eta \gamma(s)}{2\pi\left[(\xi - s)^2 + \eta^2\right]} ds \\ dV_\xi = -\dfrac{((\xi - s))\gamma(s)}{2\pi\left[(\xi - s)^2 + \eta^2\right]} d \end{cases} \tag{9}$$

A linear vortex strength distribution on the panel *j* extending from $\xi = 0$ to $\xi = l_j$ has the expression

$$\gamma(s) = \frac{\gamma_{j+1} - \gamma_j}{l_j} s + \gamma_j \tag{10}$$

where

$$l_j = \sqrt{(x_{j+1} - x_j)^2 + (y_{j+1} - y_j)^2} \tag{11}$$

The matrix of the velocities V_ξ and V_η is

$$\begin{bmatrix} V_\xi \\ V_\eta \end{bmatrix} = \frac{1}{2\pi l_j} \begin{bmatrix} (l_j - \xi)B + \eta A & \xi B - \eta A \\ -l_j - (l_j - \xi)A + \eta B & l_j - \xi A - \varepsilon B \end{bmatrix} \cdot \begin{bmatrix} \gamma_j \\ \gamma_{j+1} \end{bmatrix} \tag{12}$$

where

$$\begin{cases} A = \dfrac{1}{2} ln \dfrac{\xi^2 + \eta^2}{(\xi - l_j)^2 + \eta^2} \\ B = arctan \dfrac{l_j - \xi}{\eta} + arctan \dfrac{\xi}{\mu} \end{cases} \tag{13}$$

In order to get the velocity induced by panel *j* at the control point of the panel *i*, the coordinates of control point must be expressed from the coordinate system (x, y) in the coordinate system (ξ, η) of panel *j*, making a rotation with angle β_j and a translation in the point (x_j, y_j) as it follows

$$\begin{cases} \sin\beta_j = \dfrac{y_{j+1}-y_j}{l_j} \\ \cos\beta_j = \dfrac{x_{j+1}-x_j}{l_j} \end{cases} \tag{14}$$

$$\begin{bmatrix} \xi_C(i) \\ \eta_C(i) \end{bmatrix} = \begin{bmatrix} \cos\beta_j & \sin\beta_j \\ -\sin\beta_j & \cos\beta_j \end{bmatrix} \cdot \begin{bmatrix} x_C(i)-x_j \\ y_C(i)-y_j \end{bmatrix} \tag{15}$$

$$\begin{bmatrix} V_x(i) \\ V_y(i) \end{bmatrix} = \begin{bmatrix} \cos\beta_j & -\sin\beta_j \\ \sin\beta_j & \cos\beta_j \end{bmatrix} \cdot \begin{bmatrix} V_\xi(i) \\ V_\eta(i) \end{bmatrix} =$$

$$= \frac{1}{2\pi l_j^2} \begin{bmatrix} x_{j+1}-x_j & -(y_{j+1}-y_j) \\ y_{j+1}-y_j & x_{j+1}-x_j \end{bmatrix} \cdot$$

$$\begin{bmatrix} (l_j-\xi_C(i))B+\eta_C(i)A & \xi B-\eta_C(i)A \\ -l_j-(l_j-\xi_C(i))A+\eta_C(i)B & l_j-\xi_C(i)A-\eta_C(i)B \end{bmatrix} \cdot \begin{bmatrix} \gamma_j \\ \gamma_{j+1} \end{bmatrix}$$

or

$$\begin{bmatrix} V_x(i) \\ V_y(i) \end{bmatrix} = \begin{bmatrix} P_{11}(j,i) & P_{12}(j,i) \\ P_{21}(j,i) & P_{22}(j,i) \end{bmatrix} \cdot \begin{bmatrix} \gamma_j \\ \gamma_{j+1} \end{bmatrix} \tag{16}$$

The velocities in the coordinate system (ξ_i, η_i) of the panel *i* are

$$\begin{bmatrix} V_\xi(i) \\ V_\eta(i) \end{bmatrix} = \begin{bmatrix} \cos\beta_i & \sin\beta_i \\ -\sin\beta_i & \cos\beta_i \end{bmatrix} \cdot \begin{bmatrix} V_x(i) \\ V_y(i) \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{x_{i+1}-x_i}{l_i} & \dfrac{y_{i+1}-y_i}{l_i} \\ -\dfrac{y_{i+1}-y_i}{l_i} & \dfrac{x_{i+1}-x_i}{l_i} \end{bmatrix} \cdot \begin{bmatrix} P_{11}(j,i) & P_{12}(j,i) \\ P_{21}(j,i) & P_{22}(j,i) \end{bmatrix} \cdot \begin{bmatrix} \gamma_i \\ \gamma_{i+1} \end{bmatrix} \tag{17}$$

The velocity $V_\eta(i)$ induced in the control point of panel *i* by panel *j* is

$$V_\eta(i) = \left(-\frac{y_{i+1}-y_i}{l_i}P_{11}(j,i) + \frac{x_{i+1}-x_i}{l_i}P_{21}(j,i)\right)\gamma_i + \left(-\frac{y_{i+1}-y_i}{l_i}P_{12}(j,i) + \frac{x_{i+1}-x_i}{l_i}P_{22}(j,i)\right)\gamma_{i+1} \tag{18}$$

The $n \times n$ airfoil coefficient matrix M is generated from the 2×2 panel coefficient matrix in airfoil coordinates, $P(i, j)$ for the velocity induced at the control point *i* by panel *j*, extending from node *j* to node $j+1$, and the *n* nodal vortex strengths, γ_1 through γ_n are then obtained by numerically solving the linear system

$$M \cdot \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \dots \\ \gamma_{n-1} \\ \gamma_n \end{bmatrix} = V_\infty \begin{bmatrix} [(y_2-y_1)\cos\alpha - (x_2-x_1)\sin\alpha]/l_1 \\ [(y_3-y_2)\cos\alpha - (x_3-x_2)\sin\alpha]/l_2 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ [(y_n-y_{n-1})\cos\alpha - (x_n-x_{n-1})\sin\alpha]/l_{n-1} \\ 0.0 \end{bmatrix}$$

Once the nodal strengths are known, the velocity and pressure at any point in space can be computed by adding the velocity induced by all $n-1$ vortex panels in the free stream velocity,

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = V_\infty \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} + \sum_{i=1}^{n-1} \begin{bmatrix} V_x(i) \\ V_y(i) \end{bmatrix} \tag{19}$$

The lift coefficient for the entire airfoil is the sum of those induced by all the $n-1$ panels,

$$C_l = \sum_{i=1}^{n-1} \frac{l_i}{c} \cdot \frac{\gamma_i + \gamma_{i+1}}{V_\infty} \tag{20}$$

V. Results

For the clean airfoil at $\alpha = 0^\circ$ the flow initially separates around 50% of the chord length and this separation causes a periodic vortex shedding in the wake of the airfoil. At $\alpha = 10^\circ$ and $\alpha = 15^\circ$ the separation bubble and the vortex structures are larger and the separation point on the suction side moves upstream with increasing the angle of attack. The separated vortices tend to merge into larger structures before being shed into the wake.

The filled cavity has a strong influence on the structure of the flow in the separation bubble. It promotes smaller-scale vortex shedding than would otherwise occur for the airfoil without a cavity at the same angle of attack.

VI. Conclusions

The section lift coefficients predicted by thin airfoil theory and panel codes are in good agreement with experimental data for low Mach numbers and small angles of attack. The airfoil with filled cavity gives good results regarding the maximum lift coefficient and the behavior of the helicopter retreating blade.

References

[1] J. G. Leishman, *Principles of Helicopter Aerodynamics*, Cambridge University Press, 2007.

[2] J. Seddon, S. Newman, "*Basic Helicopter Aerodynamics*", AIAA Education Series, Reston Virginia, USA, 2001.

[3] C. Rotaru, "Nonlinear Characteristics of Helicopter Rotor Blade Airfoils: An Analytical Evaluation," *Mathematical Problems in Engineering*, vol. 2013, Article ID 503858, 9 pages, 2013.

[4] R. Prouty, Helicopter Performance, Stability and Control, Krieger Publishing Company, Florida, USA, 2002.

[5] J. Katz, A. Plotkin, "Low Speed Aerodynamics" Cambridge University Press, 2010.

[6] C. Rotaru, I. Circiu, M. Boscoianu, "Computational Methods for the Aerodynamic Design", Review of the Air Force Academy, No 2(17)/2010, p. 43-48.

[7] D. Sakhr and V. Horak "Effect of Free-stream Turbulence Properties on Boundary Layer Laminar-Turbulent Transition: A new Approach, *9th International Conference Mathematical Problems in Engineering, Aerospace and Sciences (ICNPAA 2012)*, Book Series AIP Conference Proceedings, Volume 1493, Pages 282-289, Published 2012.

[8] C. Rotaru, A. Arghiropol, "Maple soft solutions for nonlifting flow over arbitrary bodies", *Proceedings of the 3rd WSEAS International Conference on FINITE DIFFERENCES-FINITE ELEMENTS-FINITE VOLUMES-BOUNDARY ELEMENTS*, ISSN 1790-2769, pp 270-274, Romania, 2010.

Rotor-Liquid-Fundament System's Dynamics

A. B. Kydyrbekuly, L. A. Khajiyeva, G. E. Ybraev

***Abstract*—** In the present paper the dynamics of the out-of-balance rotor is examined, which is fractionally filled with an ideal fluid, placed on frictionless bearings, taking into account the elasticity of the fundament. Equations of motion of the rotor system with a concavity fractionally filled with a fluid are gained and solved. Also examinations of forced and natural oscillations of the fundament and rotor are made. Values of frequencies and amplitudes are gained and analysed.

***Keywords*—** rotor oscillations, the fundament, frictionless bearing, concavity with a fluid.

I. Introduction

The paper is devoted to development and analysis of the generalized models of dynamic systems with concavities, fractionally filled with fluids. Creation of machines at a qualitatively new level supposes the use of major achievements of fundamental sciences. In execution of these requirements an important role is assigned to high-speed mechanisms and machines. Rotor systems, which are widely used now in many industries, belong to them. In a major variety of rotor systems a considerable proportion is occupied by rotor systems with concavities containing fluid. The analysis of experimental operations displays that under certain conditions the fluid being in a concavity of the rotor, is the main source of origin of unstable modes of the system driving. In this regard the solution of problems of the dynamic analysis of rotor systems with concavities which are fractionally filled with fluid is actual and represents great practical interest. Now a majority of the rotor machines, which are widely used in industry and engineering, rotate on frictionless bearings [1]. Blocks with liquid or gas lubrication, though do have some advantages over frictionless bearings, do not find wide application in some fields of industry and engineering. The reason for that is the operation of a lubricant stratum, calling the origin of auto-oscillations in the rotor system, which are accompanied by considerable amplitudes and reduce in fast outage of bearing assemblies. At projection and estimate of vibrational performances of rotor machines it is necessary to consider housing oscillations, i.e. to consider the dynamic system "rotor-housing-fundament" as a whole [2].

In many theoretical and practical examinations on dynamics of the rotor systems containing fluid, only rotor oscillations with fluid are considered, and thus the bed (fundament) is considered immobile. Such assumption leads to essential errors at an estimate of dynamic and motion characteristics of the rotor system as a whole. Pilot studies of such dynamic systems as rotor systems display an importance of accounting of fundamental vibrations and the need of development of standards on their lowering [3].

II. Problem Statement

Let us consider the dynamics of the rotor system with concavity which is fractionally filled with an ideal fluid. The rotor has a cylindrical concavity of radius R. The rotor is placed on frictionless bearings which have a nonlinear response [4].

In this case equations of motion of the system look like

$$\left.\begin{aligned} &m\ddot{x} + 2c_0(x-x_2) + 2c_1(x-x_2)^3 + \chi\dot{x} = me\Omega_0{}^2\cos\Omega_0 t + F_x,\\ &m\ddot{y} + 2c_0(y-y_2) + 2c_1(y-y_2) + \chi\dot{y} = me\Omega_0{}^2\sin\Omega_0 t + F_y,\\ &M\ddot{x}_2 + 2c_2x_2 - 2c_0(x-x_2) - 2c_1(x-x_2)^3 + \chi_0\dot{x}_2 = 0,\\ &M\ddot{y}_2 + 2c_2y_2 - 2c_0(y-y_2) - 2c_1(y-y_2)^3 + \chi_0\dot{y}_2 = 0. \end{aligned}\right\} \quad (1)$$

$$F_x = Rh\int_0^{2\pi} P|_{r=R}\cos(\Omega_0 t+\varphi)d\phi, \quad (2)$$

$$F_y = Rh\int_0^{2\pi} P|_{r=R}\sin(\Omega_0 t+\varphi)d\phi, \quad (3)$$

where h – is the height of the concavity of the rotor, $P|_{r=R}$ – is the fluid pressure on the rotor wall.

Equations of motion of fluid look like [5]

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - 2\Omega_0\upsilon = -\frac{1}{\rho}\frac{\partial P}{\partial r} - \ddot{x}\cos(\Omega_0 t+\varphi) - \ddot{y}\sin(\Omega_0 t+\varphi),\\ &\frac{\partial \upsilon}{\partial t} + 2\Omega_0 u = -\frac{1}{\rho r}\frac{\partial P}{\partial \varphi} + \ddot{x}\sin(\Omega_0 t+\varphi) - \ddot{y}\sin(\Omega_0 t+\varphi). \end{aligned}\right\} \quad (4)$$

Equation of continuity at $\rho = const$

$$\frac{\partial(ur)}{\partial r} + \frac{\partial \upsilon}{\partial \varphi} = 0. \quad (5)$$

Boundary conditions of a hydrodynamic problem look like:

1) a requirement of equality to zero of the radial component of velocity of a particle of fluid on the rotor wall

$$u|_{r=R} = 0; \quad (6)$$

2) a requirement of equality to zero of pressure on the free surface of fluid

$$P|_{r=r_0} = 0 \text{ or } \frac{\partial P}{\partial t} = \rho\Omega_0{}^2 r_0 u|_{r=r_0}, \quad (7)$$

where u and υ – are the radial and tangential components of velocity of a particle of fluid, P and ρ – are the pressure and denseness of fluid, r_0 – is the radius of the free surface of fluid.

III. The Solution of Equations of System Motion

The solution of hydrodynamical equations (4) - (7) can be realized with several methods. Let us present component velocities of a particle of fluid, using a flow function Φ and a potential of velocities ψ, in an aspect:

$$u = \frac{\partial \Phi}{\partial r}, \quad \upsilon = \frac{1}{r}\frac{\partial \Phi}{\partial \varphi} \text{ or } u = \frac{1}{r}\frac{\partial \psi}{\partial \phi}, \quad \upsilon = -\frac{\partial \psi}{\partial r}. \quad (8)$$

Taking into account the last expressions for u and υ, we gain

$$grad\{\frac{\partial \Phi}{\partial t} + 2\Omega_0\psi + \frac{P}{\rho} + r[\ddot{x}\cos(\Omega_0 t + \phi) + \ddot{y}\sin(\Omega_0 t + \phi)]\} = 0. \quad (9)$$

The equation of continuity becomes:

$$\Delta\Phi = 0, \quad (10)$$

Δ – is a laplacian in a cylindrical frame.

From the equation (9) it is possible to define pressure

$$P = -\rho\left\{\frac{\partial \Phi}{\partial t} + 2\Omega_0\psi + r[\ddot{x}\cos(\Omega_0 t + \phi) + \ddot{y}\sin(\Omega_0 t + \phi)]\right\}. \quad (11)$$

Apparently from the equation (10), function Φ is harmonic. Introducing complex variables

$$x + iy = z, \quad x_2 + iy_2 = z_2. \quad (12)$$

let us rewrite the equation (11) in an aspect

$$P = -\rho\left[\frac{\partial \Phi}{\partial t} + 2\Omega_0\psi + \ddot{z}\exp(-i(\Omega_0 t + \varphi))\right]. \quad (13)$$

Let us consider that the rotor and the fundament make simple harmonic oscillations.

Then it is possible to present complex variables z and z_2 in an aspect

$$z = A\exp(-i\Omega_0 t) + B\exp(i\omega t), \quad (14)$$

$$z_2 = C\exp(-i\Omega_0 t) + D\exp(i\omega t). \quad (15)$$

Function Φ, as it is harmonic in correspondence with (10) and function ψ taking into account (14), is possible to be presented in an aspect

$$\Phi = R(r)\exp(i(\sigma t - \varphi)), \quad (16)$$

$$\psi = R_1(r)\exp(i(\sigma t - \varphi)). \quad (17)$$

Substituting the expression for function Φ in the formula (10), we gain an expression for function $R(r)$

$$R(r) = C_1 r + \frac{C_2}{r}. \quad (18)$$

Taking into account (8), we have an expression for function ψ

$$R_1(r) = i\left(C_1 r - \frac{C_2}{r}\right). \quad (19)$$

Using boundary conditions (6) and (7), we discover C_1 and C_2:

$$C_1 = -\frac{i\sigma\omega^2}{\left(q^2 - 1\right)\left(\gamma\sigma^2 - 2\Omega_0\sigma - \Omega_0^2\right)}, \quad (20)$$

$$C_2 = -\frac{R^2 i\sigma\omega^2}{\left(q^2 - 1\right)\left(\gamma\sigma^2 - 2\Omega_0\sigma - \Omega_0^2\right)}, \quad (21)$$

$$\sigma = \omega - \Omega_0, \quad q = \frac{R}{z_0}, \quad \gamma = \frac{q^2 + 1}{q^2 - 1}. \quad (22)$$

Substituting the discovered values C_1 and C_2 in the expressions of functions Φ and ψ, defined by formulas (16) and (17) and considering (13), from (2) and (3), we gain an expression of hydrodynamic force F_r in an aspect

$$F_r = F_x + iF_y = Am_\alpha\Omega_0{}^2\exp(i\Omega_0 t) + \\ + Bm_\alpha\omega^2\frac{(\sigma^2 - 2\Omega_0\sigma - \Omega_0{}^2)}{(\gamma\sigma^2 - 2\Omega_0\sigma - \Omega_0{}^2)}\exp(i\omega t), \quad (23)$$

$m_\alpha = \pi\rho R^2 h$ – is the mass of fluid, necessary for complete filling of the rotor concavity.

Taking into account the fact that the first and third equations of the system (1) do not depend on each other, and also the fact that rotor and fundament motions in the directions of *Ox* and *Oy* axes are identical, equations of motion of the rotor and the fundament (1) become:

ISBN: 978-1-61804-253-8

$$\left.\begin{aligned}\ddot{z}+k_0{}^2(z-z_2)+k_1(z-z_2)^3+2n\dot{z}=e\Omega_0{}^2exp(i\Omega_0 t)+\frac{F_r}{m},\\ \ddot{z}_2+k_2{}^2z_2-k_{01}{}^2(z-z_2)-k_{10}(z-z_2)^3+2n_0\dot{z}_2=0.\end{aligned}\right\} \quad (24)$$

Here $\mu=\frac{m}{M}$ – is a ratio of the rotor mass to the fundament mass,

$$k_0{}^2=\frac{2c_0}{m},\ k_1=\frac{2c_1}{m},\ 2n=\frac{k}{m},\ k_2{}^2=\frac{2c_2}{m},k_{01}{}^2=\frac{2c_0}{M}=\mu k_0{}^2,$$
$$k_{10}=\frac{2c_1}{M}=\mu k_1,2n_0=\frac{k_0}{M}.$$

Having substituted (23) in (24) and considering the fact that we gain the system of algebraic equations regarding the unknown *A*, *B*, *C*, and *D*,

We have

$$C=P_0+iP_1+A(P_2+iP_3), \quad (25)$$

where

$$P_0=\frac{\mu e\Omega_0{}^2(k_2{}^2-\Omega_0{}^2)}{m_0},\quad P_1=-\frac{2n_0\Omega_0^3}{m_0}\mu e,$$
$$P_2=\frac{\mu[(1+\mu_\alpha)\Omega_0{}^2(k_2{}^2-\Omega_0{}^2)-4nn_0\Omega_0{}^2]}{m_0},$$
$$P_3=-\frac{2\mu[(1+\mu_\alpha)\Omega_0{}^2n_0+(k_2{}^2-\Omega_0{}^2)n]\Omega_0}{m_0},$$

$\mu_\alpha=\frac{m_\alpha}{m}$ – is a ratio of the mass of fluid, necessary for complete filling of the rotor concavity, to the rotor mass;

$$P_4=\frac{\mu\omega^2(k_2{}^2-\omega^2)\left(1+\mu_\alpha\frac{\sigma^2-2\Omega_0\sigma-\Omega_0{}^2}{\gamma\sigma^2-2\Omega_0\sigma-\Omega_0{}^2}\right)-4nn_0\omega^2}{m_1},$$
$$P_5=\frac{2\mu(\gamma\sigma^2-2\Omega_0\sigma-\Omega_0{}^2+\mu_\alpha(\sigma^2-2\Omega_0\sigma-\Omega_0{}^2)n_0\omega^3)}{m_1(\gamma\sigma^2-2\Omega_0\sigma-\Omega_0{}^2)}+$$
$$+\frac{2\mu(2n\mu\omega(k_2{}^2-\omega^2)(\gamma\sigma^2-2\Omega_0\sigma-\Omega_0{}^2))}{m_1(\gamma\sigma^2-2\Omega_0\sigma-\Omega_0{}^2)},$$
$$m_1=(k_2{}^2-\omega^2)^2+4n_0{}^2\omega^2,$$
$$m_0=(k_2{}^2-\Omega_0{}^2)^2+4n_0{}^2\Omega_0{}^2,.$$
$$D=(P_4+iP_5)B. \quad (26)$$

Now using (25) and (26) it is possible to record the algebraic equations concerning *A* and *B*:

$$\left.\begin{aligned}&\frac{3k_{10}P_{75}}{2}\left(P_{23}{}^2A^2-2P_{01}P_{23}A\right)+e_2P_{45}-k_{01}^2P_{75}+\frac{3k_{10}}{2}P_{01}{}^2P_{75}+\\&+\frac{3k_{10}}{4}P_{75}{}^3B^2=0,\\&\frac{3}{4}k_{10}P_{23}{}^3A^3-\frac{9}{4}k_{10}P_{01}P_{23}{}^2A^2+[P_{23}(k_{01}^2+\frac{9}{4}P_{01}{}^2k_{10})-e_0P_{23}{}^*]A-\\&-(e_0+k_{01}{}^2+\frac{3}{4}k_{10}P_{01}{}^2)P_{01}+\frac{3}{2}P_{75}{}^2k_{10}(P_{23}A-P_{01})B^2=0.\end{aligned}\right\} \quad (27)$$

Defining from the first equation of the system (27) value B^2 and substituting it in the second equation of the system (27) we gain the equation of the third degree concerning *A*, which is easily solved with Kardano's method. After determination of *A* and *B* values from the system solution (27), further on using the equations (25) and (26) it is possible to define the stationary values *C* and *D*.

In the equations of the system (27) the labels are introduced:

$$P_{01}=P_0+iP_1,\ P_{23}=1-P_2-iP_3,\ P_{45}=P_4+iP_5,\ P_{75}=1-P_{45},$$
$$P_{23}{}^*=P_2+iP_3,\ e_0=k_2{}^2-\Omega_0{}^2+2in_0\Omega_0,\ e_2=k_2{}^2-\omega^2+2n_0i\omega.$$

The equation concerning *A* looks like

$$A^3+rA^2+sA+t=0,$$

which by means of substitution $A=y-\frac{r}{3}$ is reduced to an aspect:

$$y^3+py+q=0,$$

where

$$y_1=u+\upsilon,\ y_2=-\frac{u+\upsilon}{2}+\frac{u-\upsilon}{2}i\sqrt{3}=\varepsilon_1u+\varepsilon_2\upsilon,$$
$$y_3=-\frac{u+\upsilon}{2}-\frac{u-\upsilon}{2}i\sqrt{3}=\varepsilon_2u+\varepsilon_1\upsilon$$
$$u=\sqrt[3]{-\frac{q}{2}+\sqrt{D_0}},\upsilon=\sqrt[3]{-\frac{q}{2}-\sqrt{D_0}}$$
$$D_0=\left(\frac{P}{3}\right)^3+\left(\frac{q}{2}\right)^2,\varepsilon_{1,2}=\left(-1\pm\frac{\sqrt{3}}{2}\right).$$

From the formulas defining the unknown *A*, *B*, *C* and *D*, it is obvious that they are interdependent. It is one of singularities of a non-linear system.

Here it is necessary to mark that the problem is solved in the conjecture that the rotor and the fundament make simple harmonic oscillations.

The problem also has a solution in a case when the concavity of a rotor is fractionally filled with a viscous fluid. If to consider that the system makes simple harmonic oscillations, nonlinearity of elastic performance of bearings of the rotor does not render on the solution of equations of motion of a viscous fluid.

The case considered here, when the concavity of the rotor is fractionally filled with an ideal fluid, is selected only for the

reason to gain more or less prime formulas for an engineering estimate of the physical sense of the process, which takes place in a non-linear system.

IV. Eigentones of a non-linear system, when the concavity of the rotor is fractionally filled with an ideal fluid

Let us consider the rotor system in a conjecture that the rotor is balanced, does not have an imbalance and we pass over to the solution of the problem of oscillations of the rotor and the fundament, when the concavity of the rotor is fractionally filled with an ideal fluid.

Equations of rotor motion without imbalance on a complex plane look like:

$$\left.\begin{aligned} \ddot{z} + k_0{}^2(z - z_2) + k_1(z - z_2)^3 + 2n\dot{z} + \frac{F_r}{m} = 0, \\ \ddot{z}_2 + k_2^2 z_2 - k_{01}{}^2(z - z_2) - k_{10}(z - z_2)^3 + 2n_0\dot{z}_2 = 0. \end{aligned}\right\} \quad (28)$$

Here the former labels are saved. An expression for F_r

$$F_r = Bm_\alpha\omega^2 \frac{\left(\sigma^2 - 2\Omega_0\sigma - \Omega_0^2\right)}{\left(\gamma\sigma^2 - 2\Omega_0\sigma - \Omega_0^2\right)}. \quad (29)$$

Let us present the solution in an aspect

$$z = B\exp(i\omega t), \quad (30)$$

$$z_2 = D\exp(i\omega t). \quad (31)$$

Substituting (30) and (31) in the system (28) we gain

$$D = P_8 B, \quad (32)$$

$$P_8 = \frac{\mu\left[\omega^2\left(1 + \mu_\alpha \frac{\sigma^2 - 2\Omega_0\sigma - \Omega_0^2}{\gamma\sigma^2 - 2\Omega_0\sigma - \Omega_0^2}\right) - 2in\omega\right]}{m_2} = P_9 + iP_{10},$$

$$P_9 = \frac{\mu\left[\omega^2\left(1 + \mu_\alpha \frac{\sigma^2 - 2\Omega_0\sigma - \Omega_0{}^2}{\gamma\sigma^2 - 2\Omega_0\sigma - \Omega_0{}^2}\right)(k_2{}^2 - \omega^2) - 4nn_0\omega^2\right]}{m_1},$$

$$P_{10} = \frac{-2n_0\omega\mu\omega^2\left(1 + \mu_\alpha \frac{\sigma^2 - 2\Omega_0\sigma - \Omega_0{}^2}{\gamma\sigma^2 - 2\Omega_0\sigma - \Omega_0{}^2}\right) - 2n\omega\mu(k_2{}^2 - \omega^2)}{m_1},$$

where m_1 and m_2 are defined by the formulas

$$m_1 = (k_2{}^2 - \omega^2)^2 + 4n_0{}^2\omega^2,$$
$$m_2 = (k_2{}^2 - \omega^2) + 2in_0\omega.$$

Now from the second equation of the system it is possible to gain

$$m_2 P_8 = k_{01}^2(1 - P_8) + \frac{3k_{10}}{4}(1 - P_8)^3 B^2.$$

From the last expression B and from (32) the magnitude D are defined:

$$B = \sqrt{\frac{4\left[m_2 P_8 - k_{01}^2(1 - P_8)\right]}{3k_{10}(1 - P_8)^3}}, \quad (33)$$

$$D = (P_9 + iP_{10})\sqrt{\frac{4\left[m_2 P_8 - k_{01}^2(1 - P_8)\right]}{3k_{10}(1 - P_8)^3}}, \quad (34)$$

$$|B| = \sqrt{(\mathrm{Re}\,B)^2 + (\mathrm{Im}B)^2}, \quad (35)$$

$$|D| = \sqrt{(\mathrm{Re}\,D)^2 + (\mathrm{Im}D)^2}. \quad (36)$$

Apparently from the formulas (33)-(36) amplitudes of eigentones of the rotor $|B|$ and the fundament $|D|$, depend on the fundamental frequency of the non-linear system ω. At the fixed values of an angular velocity of the rotor Ω_0, changing smoothly the value of the fundamental frequency ω it is possible to discover the graph of association of amplitude B and D with ω and Ω_0, i.e. to gain associations of amplitude of eigentones with the fundamental frequency ω.

V. The analysis of the gained outcomes

Let us analyse the amplitudes of eigentones of the rotor and the fundament.

a) Case $n = n_0 = 0$.

If $k_2{}^2 - \omega^2 \Rightarrow 0$, i.e. when $\omega \Rightarrow \pm k_2 = \pm\sqrt{\frac{C_2}{M}}$, and also $\gamma\omega^2 - 2(\gamma+1)\Omega_0\omega + (\gamma+1)\Omega_0{}^2 \Rightarrow 0$, that is

$$\omega_{1,2} \Rightarrow \Omega_0 \frac{\gamma+1}{\gamma}\left(1 \pm \sqrt{\frac{1}{\gamma+1}}\right),$$

then we gain that $P_8 \to 0$. Then from (33) and (34) *B=0*, it follows that $D = 2\frac{\sqrt{3}}{3}\sqrt{\frac{c_0}{c_1}}\,i$, i.e. oscillations of the rotor miss, the fundament moves in the *Oy* axis direction.

If

$$(\gamma + \mu_\alpha)\omega^2 - 2(\gamma + 1 + 2\mu_\alpha)\Omega_0\omega + (\gamma + 1 + 2\mu_\alpha)\Omega_0{}^2 = 0,$$

i.e. at

$$\omega_{1,2} = \frac{(\gamma + 1 + 2\mu_\alpha)\Omega_0}{(\gamma + \mu_\alpha)}\left(1 \pm \sqrt{\frac{1 + \mu_\alpha}{\gamma + 1 + 2\mu_\alpha}}\right),$$

then $P_8 = 0$. In this case amplitudes of eigentones of the rotor

and the fundament equal $B = 2\frac{\sqrt{3}}{3}\sqrt{\frac{c_0}{c_1}}\,i$ and *D=0*, i.e. the rotor moves in the *Oy* axis direction, and the fundament is immobile.

б) Case $n \neq 0, \quad n_0 \neq 0$.

If $k_2{}^2 - \omega^2 \Rightarrow 0$, i.e. when $\omega \Rightarrow \pm k_2 = \pm\sqrt{\frac{c_2}{M}}$, and also at $(\gamma + \mu_\alpha)\omega^2 - 2(\gamma + 1 + 2\mu_\alpha)\Omega_0\omega + (\gamma + 1 + 2\mu_\alpha)\Omega_0{}^2 = 0$, or

$$\omega_{1,2} = \frac{(\gamma + 1 + 2\mu_\alpha)\Omega_0}{(\gamma + \mu_\alpha)}\left(1 \pm \sqrt{\frac{1 + \mu_\alpha}{\gamma + 1 + 2\mu_\alpha}}\right),$$

then we gain that $P_8 = -\frac{\mu n}{n_0}$. Then from (33) and (34) it follows that the amplitude of vibrations of the rotor is restricted.

If

$$\gamma\omega^2 - 2(\gamma + 1)\Omega_0\omega + (\gamma + 1)\Omega_0{}^2 \Rightarrow 0,$$

or when $\omega_{1,2} \Rightarrow \Omega_0\frac{\gamma + 1}{\gamma}\left(1 \pm \sqrt{\frac{1}{\gamma + 1}}\right)$, then $P_8 \to 0$ and the amplitude of eigentones of the rotor and the fundament accept the values *B=0*, $D = 2\frac{\sqrt{3}}{3}\sqrt{\frac{c_0}{c_1}}\,i$.

Apparently, formulas defining the coefficients $P_0, P_1, P_2, P_3, P_4, P_5, P_7$ go into expressions of amplitudes of vibrations *A*, *B*, *C*, and *D*.

VI. Conclusion

On the basis of the above stated it is possible to conclude that by definitely selecting rigidities of the leg of the fundament and its mass, and also the degree of filling of the rotor concavity, it is possible to achieve an essential diminution of the amplitude of forced vibrations and auto-oscillations (eigentones) of the rotor and the fundament. Thus the elastic installation of the fundament has that advantage which allows to introduce exterior damping necessary for deriving of better performances, than at the rigid installation, and presumes to reach much higher operation speeds, rather than the speeds which are admissible at the terrain clearancely rigidly placed fundament.

References

[1] Kydyrbekuly A.B. Research of Dynamics of a Vertical Symmetric Out-of- Balance Rotor with the Cavity Partly Filled with Liquid and Established on Elastic Foundation, in View of Nonlinearity of Elasticity of Support and Foundation Bed. – Lisbon, Portugal, september 9-12, 2013. – P.354.

[2] Gasch R., Maurer J., Sarfeld W. Soil influence on unbalance response and stability of a simple rotor-foundation system // Journal of Sound and Vibration. – 1984. – Vol.9(34). – P. 549-566.

[3] Feng N., Hahn E. Experimental identification of the pedestal in a rotor-bearing-pedestal system // Proc. 5th IFToMM International Conference on Rotor Dynamics. – Darmstadt, Germany, 1998. – P. 734-745.

[4] Rahimov E.R., Rahmatullaev A.Sh., Kydyrbekuly A.B. Dynamics of unbalanced rotor in the interaction with the other physical fields // Trans. Seventh World Congress on the theory of machines and mechanisms.– Sevilla, Spain, 1987.– P. 532-539.

[5] Kydyrbekuly A.B. Rotor-Liquid-Fundament System's Oscillation. Advances in Mechanisms Design. Mechanisms and Machine Science 8, Springer Science+Business Media Dordrecht, 2012. – P.223-229.

Conditions for the solvability and nosolvability of multivariate nonlinear filtering problems in inhomogeneous media

M. Aripov and Z. Rakhmonov

***Abstract*—** In this paper we study the global solvability and nosolvability of a nonlinear problem of non-Newtonian filtration with nonlocal boundary condition in the case of fast diffusion, which generalized early known results other authors. On the basis of the method of standard equations, self – similar analysis, the comparison method solutions the conditions of global existence and nonexistence solutions of the nonlinear filtering problem in an inhomogeneous medium found and shows the effect inhomogeneity's of the medium in these conditions. Establish the critical global existence exponent and critical Fujita exponent, which play an important role in the study of qualitative properties of nonlinear models of reaction – diffusion, thermal conductivity, filtering, and other physical, chemical, and biological processes. In the case of the global solvability the leading term of the asymptotics of solutions obtain. Using the asymptotic formula as the initial approximation for the iterative process, a numerical calculation carried out and analyzed the results. Results of numerical experiments show that the obtained results are in good agreement with the physics of the process of nonlinear filtering.

***Keywords*—** blow-up; filtration; nonlinear boundary condition, asymptotic, numerical analysis.

Consider the following nonlocal problem of non-Newtonian filtration

$$\rho(x)u_t = \nabla\left(\left|\nabla u^m\right|^{p-2}\nabla u^m\right),\quad (x,t)\in R_+^N\times(0,+\infty), \quad (1)$$

$$-\left|\nabla u^m\right|^{p-2}\frac{\partial u^m}{\partial x_1}(0,t) = u^q(0,t),\quad t>0, \quad (2)$$

$$u(x,0) = u_0(x)\ge 0,\quad x\in R_+^N, \quad (3)$$

where $\rho(x)=(1+|x|)^n$, $n>-p$, $m>1$, $q>0$, $p>1+1/m$ - are given numerical parameters, $u_0(x)$ are continuous, nonnegative bounded functions and $u_0(x)\neq 0$.

Equation (1) and its *N*-dimensional version arise in some physical models such as population dynamics, chemical reactions, heat transfer, etc. In particular, equation (1) may be used to describe the nonstationary flow in a porous medium of fluids with a power dependence of the tangential stress on the velocity of displacement under polytropic conditions [1-6]. In this case, equation (1) is called the non-Newtonian polytropic filtration equation, which has been intensively studied since the last century (see [2-13] and references therein).

In [11], the first time it was found that when $N=1$, $m=1$, $n=0$ solution of problem (1) - (3) become blow-up in finite time, if $2(p-1)/p<q<2(p-1)$. And also in this paper have been proved the following statement:

- If $0<q\le 2(p-1)/p$, then the solution of the problem (1) - (3) exists global;
- Let $q>2(p-1)$, then the solution of the problem (1) - (3) is a global, if $u_0(x)$ is sufficiently small.

Huang, Yin and Wang [4] studied the porous media equation into multi-dimensional case

$$\begin{cases} u_t = \Delta u^m,\ x\in R_+^N,\ 0<t<T,\\ -\dfrac{\partial u^m}{\partial x_1} = u^q(x,t),\ x_1=0,\ 0<t<T,\\ u(x,0)=u_0(x),\quad x\in R_+^N. \end{cases}$$

They obtain that $q_0=(m+1)/2$ and $q_c=m+1/N$.

As for equation (1) with slow diffusion $(p>2)$, Wanjuan Du and Zhongping Li [9] considered the case $m=1$, $n=0$ and obtained the critical global existence exponent $q_0=2(p-1)/p$ and the critical Fujita exponent $q_c=(1+1/N)(p-1)$.

The authors of [13] have studied the problem (1) - (3) in the fast diffusive case, when $N=1$. They obtained the critical exponent of the global existence of solutions $q_0=\dfrac{(m(n+1)+1)(p-1)}{p+n}$ and the critical Fujita exponent

M.M. Aripov is with the National University of Uzbekistan, Tashkent, 100174 Uzbekistan (corresponding author to provide phone: +998946265232; e-mail: mirsaidaripov@mail.ru).
Z. R. Rakhmonov is with the National University of Uzbekistan, Tashkent, 100174 Uzbekistan (e-mail: zraxmonov@inbox.ru.).

$q_c = m(p-1) + \frac{p-1}{n+1}$ by constructing sub and supper solutions.

This work is devoted to the study of the conditions of solvability or nosolvability of solutions of the problem (1) - (3) and the role of influence of the density distribution under these conditions on the basis of the self-similar analysis and the method of standard equations [2], to obtain the leading term global solutions with compact support problem (1) - (3) which enabled carrying out numerical experiment. Below are the main results of the study.

We introduce the notation

$$q_0 = \frac{\left(m(n+1)+1\right)(p-1)}{p+n},\ q_c = m(p-1) + \frac{p-1}{N+n}.$$

Theorem 1. *If* $0 \le q \le q_0$, *then each solution of problem (1)-(3) exists globally.*

Proof. Let

$$u_+(x,t) = e^{Lt} g(\xi)\ ,\ g(\xi) = \left(K + e^{-M\xi}\right)^{1/m},\ \xi = (1+x_1)e^{Jt},$$

$$x_i = 0,\ i = \overline{2,N},$$

where $L = J(p+n)/\left[1 - m(p-1)\right]$, $J = (K+1)^2$, $M = (K+1)^{q/[m(p-1)]}$. It is easy to calculate that

$$-\left|\frac{\partial u_+^m}{\partial x}\right|^{p-2} \frac{\partial u_+^m}{\partial x}\bigg|_{x_1=0} = -e^{(p-1)(Lm+J)t}\left|\left(g^m\right)'\right|^{p-2}\left(g^m\right)'(0),$$

$$\frac{\partial}{\partial x}\left(\left|\frac{\partial u_+^m}{\partial x}\right|^{p-2}\frac{\partial u_+^m}{\partial x}\right)(x,t) = e^{\left(Lm(p-1)+J(p+n)\right)t}\left|\left(g^m\right)'\right|^{p-2}\left(g^m\right)'(\xi),$$

$$\rho(x)\frac{\partial u_+}{\partial t}(x,t) = e^{(L-Jn)t}\xi^n\left(Lg(\xi) + J\xi g'(\xi)\right).$$

Hence we see that $(p-1)(Lm+J) = L - Jn$ and $(Lm+J)(p-1) \ge q$, and therefore, if the

$$\left(\left|\left(g^m\right)'\right|^{p-2}\left(g^m\right)'\right)'(\xi) - J\xi^{n+1}g'(\xi) - L\xi^n g(\xi) \le 0, \qquad (4)$$

$$-\left|\left(g^m\right)'\right|^{p-2}\left(g^m\right)'(0) \ge g^q(0), \qquad (5)$$

then we obtain the following readily verifiable inequality

$$\rho(x)\frac{\partial u_+}{\partial t} \ge \nabla\left(\left|\nabla u_+^m\right|^{p-2}\nabla u_+^m\right),$$

$$-\left|\nabla u_+^m\right|^{p-2}\nabla u_+^m(0,t) \ge u_+^q(0,t).$$

Not difficult to verify that if $K \ge \|u_0\|_\infty^m$ is large enough, then (4) and (5) are satisfied. Hence we have $u_+(x,0) \ge u_0(x)$ and $u_+(0,0) > u_0(0)$, by the comparison principle solutions arrive at the statement of Theorem 1.

Theorem 2. *If* $q > q_c$, *then the problem (1)–(3) admits nontrivial global solutions with small initial data.*

Proof. Equation (1) can be in the area of $Q_\infty = \left\{(x,t):\ x \in R_+^N, 0 < t < +\infty\right\}$ self-similar solution of the form

$$u_+(t,x) = (T+t)^{-\gamma}\ f(\xi), \qquad (6)$$

where $\xi = |\zeta|$, $\zeta_i = (1+x_i)(T+t)^{-\sigma}$, $i = 1,\ldots,N$,

$$\gamma = \frac{p-1}{q(p+n) - (p-1)\left(m(n+1)+1\right)},$$

$$\sigma = \frac{q - m(p-1)}{q(p+n) - (p-1)\left(m(n+1)+1\right)}.$$

We construct an supersolution of problem (1)-(3). To $u_+(t,x)$ was an supper solution of problem (1) - (3), the function $f(\xi)$ must satisfy the following inequalities [6, 7]

$$\xi^{1-N}\frac{d}{d\xi}\left(\xi^{N-1}\left|\frac{df^m}{d\xi}\right|^{p-2}\frac{df^m}{d\xi}\right) + \sigma\xi^{n+1}\frac{df}{d\xi} + \gamma\xi^n f \le 0, \quad (7)$$

$$-\left|\left(f^m\right)'\right|^{p-2}\left(f^m\right)'(1) \ge f^q(1). \qquad (8)$$

Consider the following function

$$\overline{f}(\xi) = \left(a + b\xi^{\frac{p+n}{p-1}}\right)_+^{-\frac{p-1}{1-m(p-1)}}, \qquad (9)$$

where $b = \frac{1-m(p-1)}{m(p+n)}\sigma^{1/(p-1)} > 0$, $a > 0$, $i_+ = \max(0, i)$.

We see that (7) is valid if

$$-\left(\sigma(N+n) - \gamma\right)\xi^n\overline{f} \le 0, \qquad (10)$$

it is easy to see that under the condition of Theorem 2, inequality (10) is always possible.

Substituting the function $\overline{f}(\xi)$ into (8) we obtain the following expression:

$$\sigma(a+b)^{-\frac{p-1}{1-m(p-1)}}\bigg|_{\xi=1} \ge (a+b)^{-\frac{q(p-1)}{1-m(p-1)}}\bigg|_{\xi=1} \qquad (11)$$

and it is valid, if

$$a + b \le \sigma^{-\frac{1-m(p-1)}{(q-1)(p-1)}}.$$

In conclusion, the self-similar solution $u_+(t,x)$ is supersolution of problem (1) - (3). By the principle of comparisons solutions follows: $u(t,x) \le u_+(t,x)$ in Q_∞, if $u_0(x)$ is sufficiently small.

We give the rest of the theorem without proof.

Theorem 3. *If $q > q_0$, then the solution of the problem (1)-(3) with appropriately large initial data blows up in a finite time.*

Theorem 4. *If $q_0 < q < q_c$, then each nontrivial solution of the problem (1)-(3) blows up in a finite time.*

Theorem 5. *Let $1 < p < 1+1/m$, when $\xi \to +\infty$ vanishing at infinity the solution of (7), (8) has the asymptotic*

$$f(\xi) \sim C\bar{f}(\xi),$$

where

$$C = \left(\sigma\left((N+n)\left(m(p-1)-1\right)+p+n\right)\right)^{1/[1-m(p-1)]}.$$

On the basis of the above qualitative studies were carried of numerical calculations. Results of numerical experiments show the rapid convergence of the iterative process at the expense of the successful choice the initial approximation. Below are some results of numerical experiments for different values of the numerical parameters. All figures in bold lines correspond to the initial approximation.

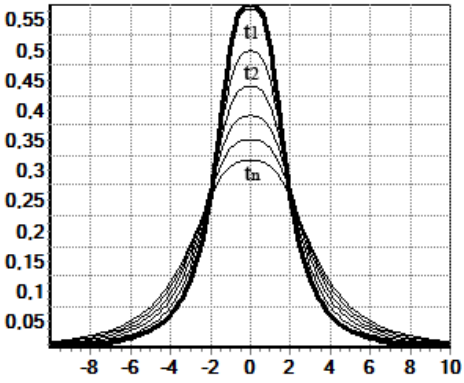


Fig 1. m=1.45, p=1.5, q=2, n=0.1

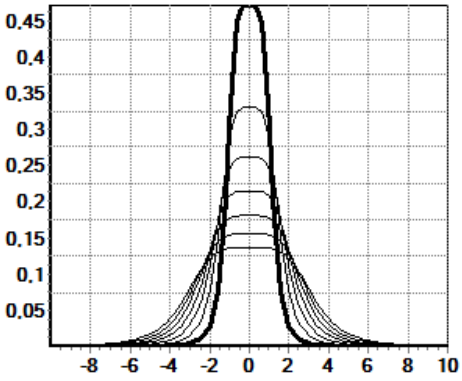


Fig 2. m=1.65, p=1.35, q=1.5, n=0.25

REFERENCES

[1]. M.Aripov, Standard Equation's Methods for Solutions to Nonlinear problems (Monograph), Tashkent, FAN, 1988.

[2]. A.S. Kalashnikov, Some problems of the qualitative theory of nonlinear degenerate second-order parabolic equations, Russian. Math. Surveys 42(2) (1987), 169-222.

[3]. *Victor A. Galaktionov and Juan L. Vazquez.* The problem of blow-up in nonlinear parabolic equations. Discrete and continuous dynamical systems, vol. 8, №2, April 2002, 399-433.

[4]. W. Huang, J. Yin, and Y. Wang, On critical Fujita exponents for the porous equation with nonlinear boundary condition, J. Math. Anal. Appl. 286 (2003), 369-377.

[5]. *Zejia W., Jingxue Y., Chunpeng W.* Critical exponents of the non-Newtonian polytropic filtration equation with nonlinear boundary condition. Appl. Math. Lett. 20, 2007, 142-147.

[6]. *Li Z., Mu Ch.* Critical exponents for a fast diffusive polytrophic filtration equation with nonlinear boundary flux. J. Math. Anal. Appl. 346, 2008, 55-64.

[7]. *Jiang Z. X. and Zheng S. N.* Doubly degenerate parabolic equation with nonlinear inner sources or boundary flux, Doctor Thesis, Dalian University of Technology, China, 2009.

[8]. *Zheng P., Mu Ch., Liu D., Yao X. and Zhou Sh.* Blow-up analysis for a quasilinear degenerate parabolic equation with strongly nonlinear source. Abstract and Appl. Anal. vol. 2012, Article ID 109546, 19 p.

[9]. *Wanjuan Du and Zhongping Li.* Critical exponents for heat conduction equation with a nonlinear Boundary condition. Int. Jour. of Math. Anal. vol. 7, 11, 2013, 517-524.

[10]. *Li Z., Mu Ch. and Du W.* Critical Fujita exponent for a fast diffusive equation with variable coefficients. Bull. Korean Math. Soc. 50. 2013, №1, 105-116.

[11]. *Galaktionov V. A., Levine H. A.* On critical Fujita exponents for heat equations with nonlinear flux boundary condition on the boundary. Israel J. Math. 94, 1996, 125-146.

[12]. Aripov M, Rakhmonov Z. Asymptotic behavior of self-similar solutions of a nonlinear problem polytropic filtration with nonlinear boundary conditions. Jour.. Comp. Tech., 2013, v.18, №4, pp.50-55.

[13]. Aripov M., Rakhmonov Z. Numerical simulation of a nonlinear problem of a fast diffusive filtration with a variable density and nonlocal boundary conditions. Proceedings of the 2014 International Conference on Mathematical Methods, Mathematical Models and Simulation in Science and Engineering, Mathematical Models and Simulation in Science and Engineering, Series 23, 2014, 72-77.

Mirsaid Aripov is professor of the Department of Informatics and Applied Programming of National University of Uzbekistan. Honored scientist of the Republic of Uzbekistan. His main research fields are Mathematical modeling of the nonlinear processes, asymptotic theory of the nonlinear ODE and PDE, computational methods of the nonlinear boundary problems, implementation of information systems, cryptology. He is member of editorial board of TWMS journal ”Pure and Applied Mathematics”, Journal of Information Security and reviewer of the Zentralblatt MATH. Members of AMS, EMS, GAMM, ISAAC.

Zafar Rakhmonov is the PhD student in National University of Uzbekistan. His research interests are in the areas of applied mathematics and mathematical physics including Mathematical modeling of the nonlinear processes, asymptotic theory of the nonlinear ODE and PDE, computational methods of the nonlinear boundary problems. He has published research articles in reputed international journals of mathematical and engineering sciences.

Discrete Nonlocal Waves

Ciprian Acatrinei

***Abstract*—A discretization scheme for field theory, in which the space time coordinates are assumed to be operators forming a noncommutative algebra, is further developed. We study here generic waves without rotational symmetry in (2+1) - dimensional scalar field theory with Heisenberg-type noncommutativity. In the representation chosen, the radial coordinate is naturally rendered discrete. Nonlocality along this coordinate, induced by noncommutativity, accounts for the angular dependence of the fields. A complete solution and interpretation of its nonlocal features are given. The exact form of standing and propagating waves on such a discrete space is found in terms of finite series. A precise correspondence is established between the degree of nonlocality and the angular momentum of a field configuration. At small distance no classical singularities appear, even at the location of the sources. At large radius one recovers the usual commutative/continuum behaviour.**

I. Introduction and Overview

Analytical calculations in relativistic quantum field theory are mostly restricted to the perturbative regime. This works well for quantum electrodynamics and weak interactions, but fails for the theory of strong interactions, where the coupling constant is not a small number. In consequence, alongside a variety of methods (semiclassical approach, toy models), discretization schemes were developed, suitable also for putting the theory on a computer [1], [2]. In the last four decades, lattice field theory became a major industry, in which remarkable results were obtained but in which important open issues still persist. Among them are the well-known fermion doubling problem and the problem of the continuum limit in the strong coupling approximation [3]. In consequence there is a need for different approaches to the discretization of field theories which could allow for progress in the two above questions.

We report here progress [5] on a different approach [6], in which discretization is obtained through assuming that the field space-time coordinates form a noncommutative algebra. Then one chooses a discrete representation of this algebra [7], obtaining a finite or denumerable number of points. The drawback is that the field now becomes nonlocal, as it depends on two points of the lattice, if no special symmetry requirements are imposed. The fact that the lattice is usually not translational invariant, and the absence of locality, allows one to bypass the Nielsen-Ninomiya theorem [4] and to put fermions on a lattice without doubling the number of modes. Concerning the continuum limit, this is best understood by realizing that the theory has become a so-called noncommmutative (NC) field theory [9], which is a relatively recent - and rare - example of a nonlocal field theory (FT) which can be handled analytically. The question is thus whether one can pass smoothly from such a field theory to a usual one. The answer seems to depend on the quantity one wishes to compute. There are for instance classical solutions which seem to become singular in the continuum limit (for a review, see [10]). On the other hand, radial waves have the correct limit [7]. Here we report on the first non-radially symmetric solution [5], in such a framework with discreteness and nonlocality explicit at the level of the degrees of freedom.

The continuum limit - which is required in order to use our methodology for usual field theories - is *well defined*. It is interesting that the degree of nonlocality of the discrete solution corresponds to the angular momentum of the continuum limit configuration. These results further suggest that an alternative numerical scheme for field theories, complementary to the current one, is a realistic goal.

The paper is organized as follows. Section 2 introduces scalar NC FT, the radial basis which discretizes the field equations of motion and the simple procedure which trades NC for nonlocality along a discrete coordinate. The nature of planar NC angular momentum is clarified and a precise connection is established between the degree of nonlocality and the angular momentum of a field configuration. In Section 3 exact solutions are derived for the discrete field equations in terms of finite series. This is the main technical part of the paper. In Section 4 sources are included - without singularities appearing. The commutative limit is taken in Section 5, clarifying the nature of the NC solutions derived previously and confirming the precise way in which nonlocality is related to nonvanishing angular momentum. The Appendix collects some of the formulae required for the technical proofs, in order to give an idea about the required mathematical manipulations.

The general notation used is quite standard: n, k represent non-negative integers, the set of which is occasionally denoted by $I\!N$; the gamma function obeys $\Gamma(z+1) = z\Gamma(z)$; the binomial coefficient is $C_n^k = \frac{n!}{(n-k)!k!}$. A discrete version of the logarithm, $H_n = 1 + 1/2 + \ldots + 1/n$, plays an important role in what follows. $\delta_{n,k} \equiv \delta(n,k)$ is the Kronecker symbol, which is one when $n = k$ and zero otherwise. For a local field ϕ, $\phi(n) \equiv \phi_n$ denotes its value at the point n. For a bilocal field Φ that depends on *two* points, $\Phi(n,n') \equiv \Phi_{n,n'}$ denotes its value when the first point is taken at n and the second at n'.

II. Equations of motion

Consider a (2+1)-dimensional scalar field ϕ, depending on space coordinates forming a Heisenberg algebra (time is commutative and remains continuous):

$$\phi(t,\hat{x},\hat{y}), \qquad [\hat{x},\hat{y}] = i\theta\hat{\mathbb{1}} \ . \tag{1}$$

Department of Theoretical Physics Horia Hulubei National Institute for Nuclear PhysicsBucharest, Romania

Manuscript received November 10, 2014

ISBN: 978-1-61804-253-8

θ is a constant with the dimensionality of an area. The scalar field ϕ is consequently a time-dependent operator acting on the Hilbert space $\mathcal{H}$ on which the algebra is represented. Since (1) implies

$$[\hat{x}, \phi(\hat{x}, \hat{y})] = i\theta \frac{\partial \phi}{\partial \hat{y}}, \qquad [\hat{y}, \phi(\hat{x}, \hat{y})] = -i\theta \frac{\partial \phi}{\partial \hat{x}},$$

the field action, written in operatorial form, reads

$$S = \int \mathrm{d}t \, \mathrm{Tr}_{\mathcal{H}} \left(\frac{1}{2}\dot{\phi}^\dagger \dot{\phi} \quad + \quad \frac{1}{2}[\hat{x}, \phi^\dagger][\hat{x}, \phi] + V(\phi^\dagger \phi) + \right. \\ \left. + \quad \frac{1}{2}[\hat{y}, \phi^\dagger][\hat{y}, \phi] \right) \tag{2}$$

allowing for classical solutions ϕ which are non-Hermitian, $\phi^\dagger \neq \phi$. We take $V(\phi^\dagger \phi) = 0$ here. The equations of motion for the field ϕ are then

$$\ddot{\phi} + \frac{1}{\theta^2}[\hat{x}, [\hat{x}, \phi]] + \frac{1}{\theta^2}[\hat{y}, [\hat{y}, \phi]] = 0. \tag{3}$$

In Cartesian coordinates, the solutions are plane waves

$$\phi \sim e^{i(k_x \hat{x} + k_y \hat{y}) - i\omega t}, \quad k_x^2 + k_y^2 = \omega^2, \tag{4}$$

formally identical to the commutative ones. In fact, due to the operators in the exponent, such waves have bilocal character [8], in agreement with the considerations to follow. If a mass term $m^2 \phi^\dagger \phi / 2$ is inserted in Eq. (2), ω^2 should be replaced by $\omega^2 - m^2$.

If the physical situation requires polar coordinates (a source emiting radiation, a circular membrane oscillating), then the harmonic oscillator, or radial, basis $\{|n\rangle\}$, $n = 0, 1, 2, \ldots$,

$$\hat{N} |n\rangle = n |n\rangle, \quad \hat{N} = \hat{a}^\dagger \hat{a}, \quad \hat{a} = \frac{1}{\sqrt{2\theta}}(\hat{x} + i\hat{y}), \tag{5}$$

is the natural one, and the equations of motion become

$$\ddot{\phi} + \frac{2}{\theta}[\hat{a}, [\hat{a}^\dagger, \phi]] = 0. \tag{6}$$

$\hat{N} = \frac{1}{2}(\frac{\hat{x}^2 + \hat{y}^2}{\theta} - 1)$ is basically (half) the radius square operator, in units of θ. Its eigenvalues n in (5) correspond to discrete points with radius growing like $\sqrt{n}$ ($n \sim \frac{r^2}{2\theta}$) for large n. The NC plane is realized via (5) as the semi-infinite discrete space of the points labeled by n.

A. Origin of the recurrence relation

We can sandwich any equation containing the operatorial field $\phi(\hat{x}, \hat{y}, t)$ between $\langle n'|$ and $|n\rangle$ states, eliminating NC in this way. The resulting field $\langle n'| \phi(t) |n\rangle \equiv \phi_{n',n}(t)$ is indeed a commutative object but also a bilocal one, as it generically depends on two different points, n and n'. Consider an harmonic time dependence $\phi_{n',n}(t) = e^{i\omega t}\phi_{n',n}$ and use $\hat{a} |n\rangle = \sqrt{n} |n-1\rangle$, $\hat{a}^\dagger |n\rangle = \sqrt{n+1} |n+1\rangle$. Eq. (6) implies that the equation of motion for $\phi_{n',n}$ is

$$\sqrt{n'+1}\sqrt{n+1}\phi_{n'+1,n+1} + \sqrt{n'}\sqrt{n}\phi_{n'-1,n-1} \\ -(n'+1+n-\lambda)\phi_{n',n} = 0. \tag{7}$$

Above, $\lambda = \frac{\theta}{2}\omega^2$. Eq. (7) is a recurrence relation of order two, describing the radial classical dynamics of a field which lives on a discrete space. The initial angular dependence of $\phi(\hat{x}, \hat{y}, t)$, due to its dependence on both $\hat{x}$ and $\hat{y}$, is not lost. It is encoded in the two-index structure of $\phi_{n',n}$, as will be seen explicitely in what follows. The operator ϕ is reconstructed from the c-numbers $\phi_{n',n}$ via

$$\phi = \sum_{n', n \in I\!N} |n'\rangle \, \phi_{n',n}(t) \, \langle n| \, . \tag{8}$$

The bilocality of the fields $\phi_{n',n}$ thus implies the nonlocality of ϕ.

B. Radial symmetry

If however the field ϕ depends only on the combination of $\hat{x}$ and $\hat{y}$ given by $\hat{N}$, $\phi = \phi(\hat{N})$, we have radial symmetry. In this case ϕ is diagonal in the $|n\rangle$ basis, $\langle n'| \phi |n\rangle = \phi_{n',n}\delta_{n',n}$, and we have a local field. Defining $\phi_{n,n} \equiv \phi_n$ for simplicity, Eq. (7) implies, for $n = 0, 1, 2, \ldots$,

$$(n+1)\phi_{n+1} + n\phi_{n-1} + (\lambda - 2n - 1)\phi_n = 0, \tag{9}$$

again with $\lambda = \omega^2\theta/2$. The expectation value $\phi_n \equiv \langle n| \phi |n\rangle$ characterizes $\phi = \sum_{n=0}^{\infty} |n\rangle \, \phi_n \, \langle n|$ uniformly at radius squared n. No angular dependence appears anymore. If a single value ϕ_{n_0} is nonzero, then $|n_0\rangle \, \phi_{n_0} \, \langle n_0|$ describes a field located at n_0.

C. A more suggestive form

Returning to the general case (8), consider the situation in which the first index of $\phi_{n',n}$ is greater than the second, $n' \geq n$. Define $\phi_n^{(m)} \equiv \phi_{n',n}$, with $m \equiv n' - n \geq 0$. The classical equation of motion of $\phi_n^{(m)}$ follows from (7),

$$\sqrt{n+m+1}\sqrt{n+1}\phi_{n+1}^{(m)} + \sqrt{n+m}\sqrt{n}\phi_{n-1}^{(m)} \\ +(\lambda - 2n - m - 1)\phi_n^{(m)} = 0. \tag{10}$$

The index m stays constant throughout the above second-order difference equation in n. To include the situation in which the second index is greater than the first, introduce $\phi_n^{(-m)} \equiv \phi_{n,n'}$, again with $m \equiv n' - n \geq 0$. $\phi_n^{(-m)}$ is easily shown to obey the same equation as $\phi_n^{(m)}$. Consequently, the independent solutions of Eq. (10) are sufficient to characterize completely both $\phi_n^{(m)}$ and $\phi_n^{(-m)}$, provided their boundary conditions are assigned independently. To summarize, if the notation

$$\phi_{n_1,n_2} \equiv \phi^{(n_1-n_2)}_{\min\{n_1,n_2\}}, \qquad m \equiv |n_1 - n_2|, \tag{11}$$

with $n_1, n_2 \in I\!N$ is adopted, the expression (8) turns into

$$\phi = \sum_{m=0}^{\infty} a_m \sum_{n=0}^{\infty} |n+m\rangle \, e^{i\omega t}\phi_n^{(m)} \, \langle n| \\ + \sum_{m=0}^{\infty} b_m \sum_{n=0}^{\infty} |n\rangle \, e^{i\omega t}\phi_n^{(-m)} \, \langle n+m| \, . \tag{12}$$

Eq. (10) implies that configurations with different m can be freely superposed in (12), as underlined by the insertion of the coefficients a_m and b_m, determined solely through initial/boundary conditions. In contrast to $|n\rangle \, \phi_n^{(0)} \, \langle n|$ which associates a value $\phi_n^{(0)}$ to the point n, $|n+m\rangle \, \phi_n^{(m)} \, \langle n|$ associates a value $\phi_n^{(m)}$ to the *two* points $n+m$ and n. Thus,

m is a measure of the delocalization of the field configuration it characterizes. Further statements can be made about the m-expansion (12) without solving the classical equation of motion (10), as we show next.

D. Angular Momentum versus Nonlocality

Ref. [8] showed that the Cartesian coordinates plane wave excitations (4) of a planar NC FT relate to one-dimensional dipoles described by a position x and an extension $\delta x = \theta p_y$, with p_y the linear momentum in the y direction. One cannot use y as a second independent variable due to $[\hat{x}, \hat{y}] \neq 0$; the conjugate variable p_y turns out to be the natural substitute.

A similar picture can be developed for radial coordinates. The bilocal quantity $\phi_n^{(m)} \equiv \phi_{n',n}$ can be described by two variables: a discrete radius squared given by n (by $2n+1$ actually) and an 'extension' $m \equiv |n'-n|$. The analogy with the plane wave scenario suggests that m is related to the quantity conjugated to the angle, i.e. to the planar angular momentum.

To prove this, we adapt Noether's theorem to the operatorial set-up of $\phi(\hat{x}, \hat{y})$ and obtain the expression for angular momentum in the $x-y$ plane, $J_z \equiv J$. First we identify the generator of rotations in the NC plane. Using

$$e^{i\alpha\hat{R}}\hat{O}e^{-i\alpha\hat{R}} = O + i\alpha[R,O] - \frac{\alpha^2}{2!}[R,[R,O]] + \cdots \quad (13)$$

and recalling $\hat{N} = \frac{1}{2}(\frac{\hat{x}^2+\hat{y}^2}{\theta} - 1) = \hat{a}^\dagger\hat{a}$, we see that

$$e^{i\alpha\hat{N}}\hat{x}e^{-i\alpha\hat{N}} = +\hat{x}\cos\alpha + \hat{y}\sin\alpha,$$
$$e^{i\alpha\hat{N}}\hat{y}e^{-i\alpha\hat{N}} = -\hat{x}\sin\alpha + \hat{y}\cos\alpha.$$

Consequently, $\hat{N}$ generates rotations in the $x-y$ plane. The variation of the field ϕ under an infinitesimal rotation is then

$$\delta\phi \equiv e^{i\alpha\hat{N}}\phi(\hat{x},\hat{y})e^{-i\alpha\hat{N}} - \phi(\hat{x},\hat{y}) \overset{\alpha\to 0}{\simeq} i\alpha[\hat{N},\phi].$$

Due to the trace over the Hilbert space appearing in (2), the field action remains invariant under such unitary transformations. One can therefore proceed and adapt the usual way of thinking of the Noether theorem to this (still classical although) operatorial set-up. The conserved charge associated to the invariance under rotations, the angular momemtum, turns out to be

$$J_z = Tr_H\ i\dot{\phi}^\dagger[\hat{a}^\dagger\hat{a},\phi]. \quad (14)$$

The commutator appearing in (14) is particularly simple in two instances,

$$\left[\hat{a}^\dagger\hat{a}, |n+m\rangle\langle n|\right] = m\ |n+m\rangle\langle n| \quad (15)$$

and

$$\left[\hat{a}^\dagger\hat{a}, |n\rangle\langle n+m|\right] = -m\ |n\rangle\langle n+m|. \quad (16)$$

These two equations already show that states 'delocalized' by an amount m are expected to carry m units of angular momentum. Using them and denoting the angular momentum of a field configuration ϕ by $J_z[\phi]$ we obtain for the expressions entering (12)

$$J_z\left[\sum_{n\in I\!N} |n+m\rangle\,\phi_n^{(m)}\,\langle n|\right] = +m\omega\ \sum_{n=0}^{\infty}[\phi_n^{(m)}]^2; \quad (17)$$

$$J_z\left[\sum_{n\in I\!N} |n\rangle\,\phi_n^{(-m)}\,\langle n+m|\right] = -m\omega\ \sum_{n=0}^{\infty}[\phi_n^{(-m)}]^2. \quad (18)$$

Dividing by the normalization factor $N_m^+ = \sum_{n=0}^{\infty}[\phi_n^{(m)}]^2$, respectively by $N_m^- = \sum_{n=0}^{\infty}[\phi_n^{(-m)}]^2$, leaves us with the results $(+m\omega)$, respectively $(-m\omega)$. No matter what the precise values of $\phi_n^{(m)}$ and $\phi_n^{(-m)}$ are, the index m determines the angular momentum, which is related exclusively to the degree of delocalization of a field configuration. The fact that N_m^+ and N_m^- may be formally infinite [similarly to the situation of commutative 2D radial waves, or even of harmonic 1D waves] is irrelevant, since the normalization factors cancel out in the final result. If N_m^+ and N_m^- are finite or properly regularized [e.g. through an upper cut-off n_{max}, or a multiplicative factor $r^{n/2} < 1$ suppressing $\phi_n^{(m)}$ at large n, to be removed only at the end of the calculations] we can further define

$$\phi^{(+m)} \equiv \frac{1}{\sqrt{N_m^+}}\sum_{n\in I\!N} |n+m\rangle\,\phi_n^{(m)}\,\langle n|, \quad (19)$$

$$\phi^{(-m)} \equiv \frac{1}{\sqrt{N_m^-}}\sum_{n\in I\!N} |n\rangle\,\phi_n^{(-m)}\,\langle n+m| \quad (20)$$

and suggestively write the general solution (12) as an m-expansion

$$\phi = \sum_{m\in I\!N}[a_m\phi^{(+m)} + b_m\phi^{(-m)}]. \quad (21)$$

The operatorial solution $\phi^{(+m)}$ has delocalization m, cf. (19), and angular momentum $+m\omega$, cf. (17). Similarly, $\phi^{(-m)}$ has delocalization $-m$, cf. (20), and angular momentum $-m\omega$, cf. (18). In consequence, (21) admits two equivalent interpretations. From the point of view of the theory defined on the discrete set of points n, it is an expansion in field configurations with well-defined nonlocality, more precisely bilocality, m. From the point of view of planar NC FT, Eq. (21) takes into account the nonradial dependence of ϕ through an angular momentum expansion.

III. Bilocal waves via finite series

In this section we solve Eq. (10) to obtain the exact form of $\phi_n^{(m)}$ and $\phi_n^{(-m)}$. The difference equation (10) describes travelling or standing waves on the semi-infinite discrete space of points $n \in I\!N$. It is convenient to parametrize its two independent solutions as follows:

$$\phi_n^{1(m)} = \sqrt{\frac{(n+m)!}{n!}}f_1(\lambda)u_n^{(m)}, \quad (22)$$

$$\phi_n^{2(m)} = \sqrt{\frac{(n+m)!}{n!}}f_2(\lambda)v_n^{(m)}. \quad (23)$$

The functions $f_1(\lambda)$ and $f_2(\lambda)$, important for normalization, will be given later. They can be temporarely neglected since they do not depend on n. $u_n^{(m)}$ and $v_n^{(m)}$, which will turn out to be two polynomials in λ, are denoted collectively by $\tilde{\phi}_n^{(m)}$. In consequence (22,23) amount for the time being to the substitution

$$\phi_n^{(m)} = \sqrt{\frac{(n+m)!}{n!}}\tilde{\phi}_n^{(m)}. \quad (24)$$

ISBN: 978-1-61804-253-8

The field $\tilde{\phi}_n^{(m)}$ then satisfies the simpler recurrence

$$(n+m+1)\tilde{\phi}_{n+1}^{(m)}+n\tilde{\phi}_{n-1}^{(m)}+(\lambda-2n-m-1)\tilde{\phi}_n^{(m)}=0. \quad (25)$$

If the discrete derivative operator Δ is defined by

$$\Delta\phi_n = \phi_{n+1} - \phi_n \quad (26)$$

and the shift operator $\hat{E}$ is defined by

$$\hat{E}\phi_n \equiv \phi_{n+1}, \quad (27)$$

then the homogeneous difference equation (25) can be rewritten as

$$\left[\hat{D}\right]\tilde{\phi}_n^{(m)} \equiv \left[n\Delta^2\hat{E}^{-1}+(m+1)\Delta+\lambda\right]\tilde{\phi}_n^{(m)}=0. \quad (28)$$

The operator $\left[\hat{D}\right]$ annihilates the field $\tilde{\phi}_n^{(m)}$, $\forall n \in I\!N$.

A. First solution

Define for α real, the falling factorial power $n^{\underline{\alpha}}$,

$$n^{\underline{\alpha}} \equiv \frac{\Gamma(n+1)}{\Gamma(n+1-\alpha)}, \qquad \stackrel{(26)}{\Longrightarrow} \qquad \Delta n^{\underline{\alpha}} = \alpha n^{\underline{\alpha-1}}. \quad (29)$$

For natural $\alpha = k$, $n^{\underline{k}} = n(n-1)\cdots(n-k+1)$; this explains the name.

We search for a solution of the form

$$\tilde{\phi}_n^{(m)} = a_0^m(\sigma,\lambda)n^{\underline{\sigma}} + a_1^m(\sigma,\lambda)n^{\underline{\sigma+1}} + a_2^m(\sigma,\lambda)n^{\underline{\sigma+2}} + \ldots, \quad (30)$$

i.e. an expansion in falling factorial powers of n. Equating to zero the coefficient of each $n^{\underline{k+\sigma}}$ in

$$\left[\hat{D}\right]\sum_{k=0}^{\infty} a_k^m(\sigma,\lambda)n^{\underline{\sigma+k}} = 0 \quad (31)$$

we obtain the indicial equation

$$\sigma(\sigma+m)=0 \quad (32)$$

and the recurrence relation for the expansion coefficients $a_k^m(\sigma,\lambda)$

$$(k+\sigma)(k+\sigma+m)\; a_k^m(\sigma,\lambda) + \lambda\; a_{k-1}^m(\sigma,\lambda) = 0. \quad (33)$$

Eq. (33) guarantees that (30) is also an expansion in powers of λ. For $\sigma \to 0$ we obtain the first finite series solution

$$u_n^{(m)} = \sum_{k=0}^{n} \frac{(-\lambda)^k}{k!(m+k)!}\left[\frac{\Gamma(n+1)}{\Gamma(n+1-k)} = n^{\underline{k}}\right]. \quad (34)$$

From now on we drop the dimensionfull multiplicative constant a_0 by putting $a_0 \equiv 1$. It can be reintroduced whenever required.

B. Second solution - infinite series

The second solution cannot be obtained easily since we confront the case of roots differing by an integer in (32); taking $\sigma = -m$ makes the coefficients diverge after some k. We therefore use the following procedure [11]: we solve the inhomogeneous equation

$$\left[\hat{D}\right]\sum_{k=0}^{\infty} a_k(\sigma)n^{\underline{\sigma+k}} = \sigma(\sigma+m)^2 n^{\underline{\sigma}}, \quad (35)$$

take the derivative of its solution with respect to σ, and then take the limit $\sigma+m \equiv \epsilon \to 0$. In the process, one obtains terms involving the digamma function $\Psi(x) = \frac{d\log\Gamma(x)}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}$. Using the poles-displaying expansion (67), then carefully observing the multiplicative cancellation between zeroes and poles in the limit $\sigma+m \to 0$ we obtain the second solution

$$\begin{aligned} w_n^{(m)} \;=\; & -n!\sum_{k=0}^{m-1}\frac{\lambda^k}{k!}\frac{(m-k-1)!}{(m-k+n)!} \\ & + \sum_{k=n+1}^{\infty}\frac{\lambda^{k+m}(-)^n n!(k-n-1)!}{k!(k+m)!} \\ & + \sum_{k=0}^{n}\frac{\lambda^m(-\lambda)^k}{(k+m)!}\,C_n^k \times \\ & \times(H_{n-k}-H_k-H_{k+m}+H_{m-1}-\gamma). \end{aligned} \quad (36)$$

As already mentioned, H_k is a discrete version of the logarithmic function,

$$H_k = 1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{k}\,, \qquad k=1,2,3\ldots \quad (37)$$

with $H_0 = 0$ however, whereas γ is the Euler-Mascheroni constant,

$$\gamma = \lim_{k=\infty}(H_k - lnk) \simeq 0.5772.$$

C. Second solution - finite series

Although $w_n^{(m)}$ solves (25), it is an infinite convergent series in λ. To obtain a finite series, we search for a linear combination of $u_n^{(m)}$ and $w_n^{(m)}$,

$$v_n^{(m)} = a(\lambda)u_n^{(m)} + b(\lambda)w_n^{(m)}, \quad (38)$$

that is independent of λ in $n=0$ and $n=1$. Choosing to impose

$$b(\lambda) = e^{-\lambda} \quad (39)$$

and determining $a(\lambda)$ from the condition

$$v_0^{(m)} = a(\lambda)u_0^{(m)} + b(\lambda)w_0^{(m)} = 0, \quad (40)$$

we obtain, after a long calculation featuring unexpected but essential simplifications [5],

$$v_1^{(m)} = a(\lambda)u_1^{(m)} + b(\lambda)w_1^{(m)} = \frac{1}{m+1}, \quad (41)$$

which is constant (and convenient). Using in (38) the values of $a(\lambda)$ and $b(\lambda)$ so determined leads to a long calculation involving products and sums of finite and infinite power series.

Making use of the identities (62-64), (65-66) and a few others given in the Appendix of [5], we arrive at

$$v_n^{(m)} = \sum_{L=0}^{n-1}(-\lambda)^L \left\{ \sum_{s=1}^{n-L} \frac{(-)^{s-1}\, C_n^{s+L}\, \Gamma(m+s)}{\Gamma(m+s+L+1)} \right\}. \quad (42)$$

This is our final result for $v_n^{(m)}$. The first $v_n^{(m)}$'s can be calculated easily. As expected, $v_0^{(m)} = 0$ for any natural m; then, $\forall m \in I\!N$,

$$v_1^{(m)} = \frac{1}{(m+1)!/m!}, \quad v_2^{(m)} = \frac{(3+m)-\lambda}{(m+2)!/m!},$$

$$v_3^{(m)} = \frac{(m^2+6m+11)-(2m+8)\lambda+\lambda^2}{(m+3)!/m!}.$$

Concerning the functions f_1 and f_2 in (22, 23), the continuum limit behaviour of $u_n^{(m)}$ and $v_n^{(m)}$ requires, cf. Eqs. (60) and (61),

$$f_1(\lambda) = \lambda^{\frac{m}{2}}, \qquad f_2(\lambda) = \lambda^{-\frac{m}{2}}. \quad (43)$$

Putting together (22), (23),(43), (34) and (42) we obtain our final result for the two finite series that solve (10):

$$\phi_n^{1(m)} = \sqrt{\lambda^m \frac{(n+m)!}{n!}} \sum_{k=0}^{n}(-\lambda)^k \left[\frac{C_n^k}{(k+m)!}\right]; \quad (44)$$

$$\phi_n^{2(m)} = \sqrt{\lambda^{-m} \frac{(n+m)!}{n!}} \sum_{L=0}^{n-1}(-\lambda)^L \times \times \left\{ \sum_{s=1}^{n-L} \frac{(-)^{s-1} C_n^{s+L}(m+s-1)!}{(m+s+L)!} \right\}. \quad (45)$$

The solutions (44,45) are *finite sums* and are linearly independent, since their Casoratian (the discrete analogue of the Wronskian)

$$D_n^m \equiv \phi_n^{1(m)}\phi_{n+1}^{2(m)} - \phi_{n+1}^{1(m)}\phi_n^{2(m)} \quad (46)$$

is nonvanishing:

$$D_n^m = \frac{1}{\sqrt{n+m+1}\sqrt{n+1}} \neq 0. \quad (47)$$

The general solution of (10) is a linear combination $\phi_n^{(m)} = c_1\phi_n^{1(m)} + c_2\phi_n^{2(m)}$, with coefficients c_1 and c_2 determined from appropriate boundary conditions. Since $\phi_n^{(-m)}$ obeys the same equation (10), its general form will also be a superposition of the solutions (44) and (45) but with coefficients determined from independently assigned boundary conditions, $\phi_n^{(-m)} = c_1'\phi_n^{1(m)} + c_2'\phi_n^{2(m)}$.

Introducing now the expressions found for $\phi_n^{(m)}$ and $\phi_n^{(-m)}$ in (12) produces the general solution of the difference equation of motion (10). As already proved in Section 3, m (respectively $-m$) characterizes simultaneously the degree of nonlocality and the angular momentum of each bilocal configuration $|n+m\rangle\, \phi_n^{(m)}\, \langle n|$ (respectively $|n\rangle\, \phi_n^{(-m)}\, \langle n+m|$) in (12). The coefficients a_m and b_m are determined by the imposed initial/boundary conditions, for instance $a_m = \delta_{m,m_0}$ and $b_m \equiv 0$ if only a field configuration of angular momentum $m\omega$ is excited.

D. Connection to Laguerre polynomials

It is worthwhile to mention a simple mathematical connection. If one writes $\phi_n^{(m)} \sim \sqrt{\frac{n!}{(n+m)!}} L_n^{(m)}$ [in contradistinction to (24)] one obtains

$$(n+1)L_{n+1}^{(m)} + (n+m)L_{n-1}^{(m)} + (\lambda - m - 2n - 1)L_n^{(m)} = 0. \quad (48)$$

This is precisely the recurrence relation for the Laguerre polynomials, and it has *two* independent solutions expressible as polynomials in λ. The first, $L_n^{1(m)} = \frac{(n+m)!}{n!} u_n^{(m)}$, gives the well-known Laguerre polynomials (70). The second finite series solution of (48) provides some kind of "Laguerre polynomials of the second kind" $L_n^{2(m)}$ [which obey the Laguerre recurrence relation in n, (48), but not the Laguerre differential equation in λ] related to our $v_n^{(m)}$ through $L_n^{2(m)} = \frac{(n+m)!}{n!} v_n^{(m)}$. Hence, we also obtained in this subsection a second finite series solution of the Laguerre recurrence relation (48), apparently unavailable in the literature.

E. Radially symmetric solutions

If one has radial symmetry, $\phi = \phi(\hat{N})$, the relevant difference equation simplifies to the local form (9). Its first solution [7] follows immediately by taking $m = 0$ in (34)

$$u_n \equiv u_n^{(0)} = \sum_{k=0}^{n} \frac{(-\lambda)^k}{k!}\, C_n^k\,, \quad (49)$$

To obtain the second solution one is confronted when $m = 0$ with an indicial equation $\sigma^2 = 0$ that has roots which are equal, not differing by an integer. In spite of slight differences in proceeding due to that aspect, the resulting infinite series solution is still the specialization of (36) to $m = 0$, namely $w_n \equiv w_n^{(0)}$.

Searching for a polynomial solution $v_n = a(\lambda)u_n + b(\lambda)w_n$ that obeys $v_0 = 0$ and $v_1 = 1$ one obtains

$$v_n = \sum_{N=0}^{n-1}(-\lambda)^N \left\{ \sum_{s=1}^{n-N} \frac{(-)^{s-1} C_n^{s+N}}{s(s+1)\cdots(s+N)} \right\} \quad (50)$$

which could have been obtained directly by taking $m = 0$ in our general result (42) . Further use of (64) allows to reach an even simpler expression

$$v_n = \sum_{k=0}^{n-1} \frac{(-\lambda)^k}{k!} \sum_{j=1}^{n-k} \frac{C_{n-j}^k}{k+j}\,. \quad (51)$$

Both forms (50) and (51) are useful: the first is handier for some analytical manipulations whereas the second permits faster numerical evaluations.

IV. Sources

Sources can be taken into account by introducing an inhomogeneous term j_n in (10),

$$\sqrt{n+m+1}\sqrt{n+1}\,\Phi_{n+1}^{(m)} + \sqrt{n+m}\sqrt{n}\,\Phi_{n-1}^{(m)} + (\lambda - 2n - m - 1)\Phi_n^{(m)} = j_n. \quad (52)$$

The solution of Eq. (52) takes into account an arbitrary distribution of sources j_n and is consequently denoted by

ISBN: 978-1-61804-253-8

$\Phi_n^{(m)}[j_n]$. It can be obtained through the linear superposition of solutions $\Phi_n^{(m)}[j\delta_{n,n_0}]$ which solve Eq. (52) with sources $j_n = j\delta_{n,n_0}$ localized at an arbitrary but single point n_0.

If $n_0 = 0$ one has the most interesting case - a source at the origin, $j_n = j\delta_{n,0}$. If $j = \sqrt{m!}\lambda^{-m/2}$, the difference equation (52) is now solved precisely by $\phi_n^{2(m)}$ ($+c\phi_n^{1(m)}$, c arbitrary). Indeed, our hardly won $\phi_n^{2(m)}$ does solve the homogeneous equation (10) everywhere except at $n = 0$, simply due to the fact that there the difference equation becomes first order and admits only one solution, which is $\phi_n^{1(m)}$. If a source $j = \sqrt{m!}\lambda^{-m/2}$ is added at $n = 0$ however, $\phi_n^{2(m)}$ is the particular solution of the resulting inhomogeneous equation. This is in line with the fact that v_n enters the description of radially propagating waves, which *require* a source at the origin. If we take j arbitrary in $j_n = j\delta_{n,0}$, the solution will be (we write $j \equiv j_0$ for later convenience)

$$\Phi_n^{(m)}[j_0\delta_{n,0}] = j_0 \frac{\lambda^{m/2}}{\sqrt{m!}}\, \phi_n^{2(m)}. \tag{53}$$

Consider then the case of a source at $n_0 \geq 1$ (a ring-like source). We search for a solution of (52) with source $j_n = j\delta_{n,n_0}$. A straightforward adaptation to the discrete case of the method of variation of constants (or of the method of Green functions) suggests to search for

$$\Phi_n^{(m)}[j\delta_{n,n_0}] = c_1(n)\phi_n^{1(m)} + c_2(n)\phi_n^{2(m)}, \tag{54}$$

with c_1 and c_2 jumping only between n_0 and $n_0 + 1$, namely

$$c_1(n+1) - c_1(n) = d_1\delta_{n_0,n},$$

$$c_2(n+1) - c_2(n) = d_2\delta_{n_0,n}.$$

Eqs. (52) and (10) imply then

$$d_1 = \frac{-j\ \phi_{n_0}^{2(m)}}{\sqrt{n_0+m+1}\sqrt{n_0+1}D_{n_0}^m},$$

$$d_2 = \frac{+j\ \phi_{n_0}^{1(m)}}{\sqrt{n_0+m+1}\sqrt{n_0+1}D_{n_0}^m},$$

where $D_{n_0}^m$ is the Casoration of $\phi_n^{1(m)}$ and $\phi_n^{2(m)}$ at $n = n_0$, cf. Eq. (46). The solution for a source $j\delta_{n,n_0}$ of intensity j and location n_0 is consequently

$$\begin{aligned}\Phi_n^{(m)}[j\delta_{n,n_0}] &= -j\theta_{n,n_0}\left(\phi_n^{1(m)}\phi_{n_0}^{2(m)} - \phi_n^{2(m)}\phi_{n_0}^{1(m)}\right) \\ &\quad c_1\phi_n^{1(m)} + c_2\phi_n^{2(m)},\end{aligned} \tag{55}$$

with the step function $\theta_{m,n} \equiv \theta(m-n)$ defined as

$$\theta_{m,n} = \begin{cases} 1 & \text{if } m-n \geq 0 \\ 0 & \text{if } m-n < 0. \end{cases} \tag{56}$$

Finally, denoting by $\Phi_n^{(m)}[0] = c_1\phi_n^{1(m)} + c_2\phi_n^{2(m)}$ the general solution in the absence of sources, we obtain the general solution with an arbitrary source distribution j_n as

$$\begin{aligned}\Phi_n^{(m)}[j_n] &= -\sum_{n_0} j_{n_0}\theta_{n,n_0}\left(\phi_n^{1(m)}\phi_{n_0}^{2(m)} - \phi_n^{2(m)}\phi_{n_0}^{1(m)}\right) \\ &\quad +\Phi_n^{(m)}[0].\end{aligned} \tag{57}$$

The sum is taken over all points n_0 with nonzero sources, $j_{n_0} \neq 0$. If a source j_0 appears at the origin one notes that, due to $\phi_0^{1(m)} = \frac{\lambda^{m/2}}{m!}$ and $\phi_0^{2(m)} = 0$, the $n_0 = 0$ contribution in (57) reproduces (53). The general solution (57) does not display singularities, even at the location of the sources.

V. Commutative limit : types of waves

It is useful to see first what happens to the difference equation (10) in the commutative limit. As $\theta \to 0$, one has $n \simeq n' \simeq \frac{r^2}{2\theta} \to \infty$ and $\lambda = \frac{\theta\omega^2}{2} \to 0$, but $\lambda \cdot n \sim (\frac{\omega r}{2})^2$ is finite; m stays finite as well. In this limit the difference operator applied in (10) to a function of n, $\phi_n^{(m)}$, becomes the m-th order Bessel operator applied to a function of r, call it $f^{(m)}(r)$. Indeed, taking $n \to \infty$, expanding the square roots in (10) to order $O(\frac{1}{n^2})$ and replacing $\frac{\Delta}{\Delta n}$ by $\frac{d}{dn}$ one obtains

$$\left(n + \frac{m}{2}\right)\frac{d^2\phi_n^m}{dn^2} + \frac{d\phi_n^m}{dn} + \left(\lambda - \frac{m^2}{4n}\right)\phi_n^m = 0. \tag{58}$$

Recalling that $\lambda = \frac{\theta\omega^2}{2}$ and passing via $n \equiv \frac{r^2}{2\theta}$ from a function of n to a function of r, $\phi_n^{(m)} \to f^{(m)}(r)$, Eq. (58) becomes as $\theta \to 0$:

$$\frac{d^2f^{(m)}}{dr^2} + \frac{1}{r}\frac{df^{(m)}}{dr} + (\omega^2 - \frac{m^2}{r^2})f^{(m)}(r) = 0. \tag{59}$$

This is precisely the Bessel equation of order m for a function of independent variable ωr; The solutions of (10) should therefore reduce at large distances to linear combinations of the cylindrical functions of order m, (68) and (69).

Our solutions are consistent with the above limit. Indeed, as $n \simeq \frac{r^2}{2\theta} \to \infty$, $\lambda = \frac{\theta\omega^2}{2} \to 0$ and $\lambda \cdot n \simeq \left(\frac{\omega r}{2}\right)^2$, Eqs. (34) and (68) imply that a properly normalized $u_n^{(m)}$ becomes, as a function of r, the m-th order Bessel function $J_m(\omega r)$,

$$\sqrt{\lambda^m \frac{(n+m)!}{n!}}\ u_n^{(m)} \longrightarrow J_m(\omega r). \tag{60}$$

In this way we also establish the function f_1 in (22) to be $\lambda^{m/2}$, up to multiplicative factors which go to 1 when λ goes to 0. For instance, ortogonality of the $\phi_n^{1,m}$ in (44), seen as functions of $\lambda \in [0, \infty)$, requires further multiplication of the LHS of (60) by $\mathrm{e}^{-\frac{\lambda}{2}}$. Not being of immediate relevance for our purposes, such factors are omitted.

For the second solution things are less immediate. It is convenient to first consider $w_n^{(m)}$. When $n \to \infty$ and $\lambda \to 0$, the middle term - the infinite sum - in (36) vanishes [5]. Second, $v_n^{(m)}$ an $w_n^{(m)}$ have the same commutative limit as $\lambda \to 0$. Using this observation and Eqs. (36), (68) and (69) we obtain

$$\sqrt{\lambda^{-m}\frac{(n+m)!}{n!}}v_n^{(m)} \longrightarrow c_m\ J_m(\omega r) + d_m\ Y_m(\omega r). \tag{61}$$

where $c_m = H_{m-1} - 2\gamma - \log\frac{\theta\omega^2}{2}$ and $d_m = \pi$. This also establishes the function f_2 in (23) to be $\lambda^{-m/2}$.

The correspondence between NC and usual waves can now be established through their behaviour at large distances. Given that the m-th order Bessel function $J_m(\omega r)$ describes usual radially standing waves (oscillations), Eq. (60) implies that its

counterpart $\phi_n^{1(m)}$ describes radially standing NC waves (oscillations). Moreover, $J_m(\omega r)$ and $Y_m(\omega r)$ can carry angular momentum ωm or $-\omega m$ in usual planar field theory, in perfect agrement with the fact that $\phi_n^{1(m)}$ enters (12) in combination with either $|n+m\rangle\langle n|$ or $|n\rangle\langle n+m|$.

On the other hand, the first Hankel function of order m

$$H_m^1(\omega r) = J_m(\omega r) + iY_m(\omega r)$$

describes waves which propagate outward radially and rotate angularly with frequency plus or minus ωm [unless waves with m and $-m$ dependence are superposed, e.g. to render the angular part of the wave standing]. In consequence, the linear combination of $\phi_n^{1(m)}$ and $\phi_n^{2(m)}$ which tends to $H_m^1(\omega r)$ as $\theta \to 0$ will describe a NC wave radially propagating outwards towards $n = \infty$ and carrying angular momentum $+m\omega$ or $-m\omega$ [unless two waves with opposite m are superposed]. This combination is easily found to be

$$\phi_n^{3(m)} = \frac{i}{\pi}\left(\phi_n^{2(m)} - \phi_n^{1(m)}\left[H_{m-1} - 2\gamma - \frac{\pi}{i} - \log\frac{\theta\omega^2}{2}\right]\right)$$

and displays angular momentum $m\omega$ when it combines with $|n+m\rangle\langle n|$ and angular momentum $-m\omega$ when it combines with $|n\rangle\langle n+m|$. Similarly, the easily found linear combination of $\phi_n^{1(m)}$ and $\phi_n^{2(m)}$ which, as $\theta \to 0$, tends to $H_m^2(\omega r) = J_m(\omega r) - iY_m(\omega r)$ will describe "radially collapsing" NC waves.

APPENDIX
USEFUL FORMULAE

We collect here some of the formulae useful for the mathematical manipulations in the paper.

Recall first that $H_n \equiv 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$. Then we have the known identity

$$\sum_{k=1}^{n} \frac{(-)^{k-1}C_n^k}{k} = H_n \tag{62}$$

together with two useful generalizations of it,

$$\sum_{k=1}^{n} \frac{(-)^{k-1}}{k}\frac{C_n^k}{C_{k+m}^k} = H_{n+m} - H_m, \tag{63}$$

$$\sum_{k=1}^{n} \frac{(-)^{k-1}}{k}\frac{C_{n+m}^{k+m}}{C_{k+p}^k} = \sum_{j=1}^{n}\frac{C_{m+n-j}^m}{p+j}. \tag{64}$$

One can translate the denominator k in (62) by a positive integer amount p, to find the identity (this time one can extend the sum to $k = 0$)

$$\sum_{k=0}^{n} \frac{(-)^{k-1}C_n^k}{k+p} = -\frac{(p-1)!n!}{(p+n)!} \qquad p \in I\!N^*. \tag{65}$$

Its generalization is, again for p integer and $p \geq 1$,

$$\sum_{k=0}^{n} \frac{(-)^{k-1}C_n^k}{(k+p)C_{k+p+m}^{k+p}} = -\frac{(p-1)!(n+m)!}{(p+n+m)!}. \tag{66}$$

A few more complex identities are required to prove that $v_n^{(m)}$ is a polynomial of order $n-1$ in λ. Those identities are given in the Appendix of [5].

The digamma function $\Psi(x) = \frac{d\log\Gamma(x)}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}$, has the following useful properties [12] (z complex but not a negative integer, n positive integer)

$$\Psi(z+1) \equiv \frac{\Gamma'(z+1)}{\Gamma(z+1)} = -\gamma + \sum_{l=1}^{\infty}\frac{z}{l(l+z)}, \tag{67}$$

in particular $\Psi(n+1) = -\gamma + H_n$. The series expansion of the Bessel functions [12] is required while taking the commutative limit of $u_n^{(m)}$,

$$J_m(z) = \sum_{k=0}^{\infty}\frac{(-)^k(z/2)^{2k+m}}{k!(k+m)!}, \tag{68}$$

while the Neumann functions Y_m [12] are also needed for $w_n^{(m)}$,

$$\begin{aligned}\pi Y_m(z) = \quad & 2J_m(z)(\gamma + \log\frac{z}{2}) \\ - \quad & \sum_{k=0}^{m-1}\frac{(m-k-1)!}{k!}\left(\frac{z}{2}\right)^{2k-m} \\ - \quad & \sum_{k=0}^{\infty}(-)^k\left(\frac{z}{2}\right)^{2k+m}\frac{H_{m+k}+H_k}{k!(k+m)!}.\end{aligned} \tag{69}$$

Finally, the Laguerre polynomials are given by [12]

$$L_n^m(\lambda) = \sum_k^n \frac{\Gamma(n+m+1)}{\Gamma(k+m+1)}\frac{(-\lambda)^k}{k!(n-k)!}. \tag{70}$$

ACKNOWLEDGEMENT

We acknowledge financial support from the CNCS through the *Idei* Program Contract Nr. 121/2011 and the *Nucleu* Program Contract PN-09370102/2009.

REFERENCES

[1] K. Wilson, *Phys. Rev.* D10, 2445 (1974).
[2] J. Kogut, L. Susskind, *Phys. Rev.* D11, 395 (1975).
[3] Y. Makeenko, *Methods in Contemporary Gauge Theory*, Cambridge University Press 2005.
[4] H.B. Nielsen, M. Ninomiya, *Nucl. Phys.* B185, 20 (1981).
[5] C. S. Acatrinei *J. High Energy Phys.* 1302, 057 (2013).
[6] Acatrinei, C.S. "A New Discretization Scheme in Field Theory" pp. 122-127 in *Mahematical Methods and Computational Science*, Gh. Adam, Ján Buša, M. Hnatič (Eds.), Springer 2012.
[7] C.S. Acatrinei, *J. Phys.* A41, 215401 (2008).
[8] C.S. Acatrinei, *Phys. Rev.* D67, 045020 (2003).
[9] M.R. Douglas and N.A. Nekrasov, *Rev. Mod. Phys.* 73, 977 (2001); R.J. Szabo, *Phys. Rep.* 378, 207 (2003).
[10] J. A. Harvey, e-Print hep-th/0102076 (2001).
[11] L.M. Milne-Thomson, *The Calculus of Finite Differences*, MacMillan 1933.
[12] M. Abramowitz and I. A. Stegun (Eds.), *Handbook of Mathematical Functions*, National Bureau of Standards 1972.

Fundamental solutions of Lame's equations for granular media.

Rozin Leonid, Zdanchuk Elizaveta

***Abstract*—** In this work we consider a model for granular medium. Reduced Cosserat continuum was suggested as possible model to describe granular materials. Reduced Cosserat Continuum is an elastic medium, where all translations and rotations are independent. Moreover the force stress tensor is asymmetric and the couple stress tensor equal to zero. Here we obtain/establish the fundamental solution of Lame's equations for the reduced Cosserat model provided that concentrated unit amplitude force, harmonically dependent on time is applied to a homogeneous, unbounded, isotropic elastic medium**.**

***Keywords*—** granular medium, Lame's equation, reduced Cosserat continuum.

I. Introduction

In this work we consider a model for granular medium. By granular medium we understand a set of contact relied solid grains. The total volume of such a medium is composed of solid grains and voids, the later could be filled either with air or with a liquid. As for all granular medium grains at the surface can easily undertake a relative movement with respect to their neighbours. However, free movement of the grains requires space around them, in other words the increase of the body volume.

This type of medium and its specific behavior is very important in different branches of engineering and industrial applications such as mining, agriculture, construction and geological processes.

In most of recent papers on this subject [1] - [4] the granular media was systematically presented as discrete medium. In this close to reality description each case should have its own model that takes into account the structure and nature of the probable distribution of forces and deformations in the medium. Contrary to the previous method the use of a continuum model for such description is more general.

Granular medium occupies an intermediate position between liquids and solids. The ability of the grain to move one relative to the others makes them look like a liquid. At the same time each particle of the granular medium, taken separately, has all the properties of a solid. In this type of medium grain's size and nearest-neighbour distance are roughly comparable, rotational degrees of freedom must be considered along with the translational. To obtain a more accurate model of a granular medium it is necessary to use the continuum model with microstructure [5]-[7]. This widely known Cosserat continuum is the subject of intensive scientific activities in recent years [8]-[10]. In this paper, we propose less known the *reduced* Cosserat continuum as a model of a granular medium. In this continuum translations and rotations are independent, stress tensor is not symmetric and couple stresses tensor is equal to zero.

Originally the idea of an equal footing for rotational and translational degrees of freedom appeared in [11]. The authors obtained good correspondence between their theoretical results and experimental data. Some complementary studies of this model were also performed in more recent works [12]-[16]. These papers discuss the propagation of waves in the bulk and on the surface of the medium [12], behavior of the medium in the case of spherical prestressing [13], nonlinear [14] and thermodynamic equations [15], [16] for the reduced Cosserat continuum.

The main goal of this work is to found translation and rotation vectors caused by arbitrarily directed concentrated force, harmonically dependent on time, which is applied to a given point of a homogeneous, unbounded, isotropic elastic medium. Some relevant applications could be found in seismology [17]. In particular one is usually interested in a time evolution of waves, observed at a fixed distance from the source. For such a problem the assumption of a point-wise volume force is close to reality description of the seismic source displacement observed during the earthquake.

II. Math

As we know from [13], the Lame's equations for reduced Cosserat continuum can be expressed as

$$(\mu+\alpha)\Delta\mathbf{u}-(\lambda+\mu-\alpha)grad\ div\mathbf{u}+2\alpha rot\varphi+\mathbf{X}=\rho\ddot{\mathbf{u}} \quad (1)$$

$$2\alpha rot\mathbf{u}-4\alpha\varphi=J\ddot{\varphi}, \quad (2)$$

where u is translation vector, φ is rotation vector, λ,μ,α are elastic constants of the reduced Cosserat continuum, J is volume density of the spherical tensor of inertia, **X** is volume force.

Let us find a solution in the form of complex functions:

$$\mathbf{u}=\mathbf{u}^{*}e^{-i\omega t}, \quad (3)$$

L. Rozin is with the Department of Structural mechanics, Saint Petersburg State Polytechnic University, Saint Petersburg, Russia (e-mail: ezlarh@mail.ru).

E. Zdanchuk is with the Department of Structural mechanics, Saint Petersburg State Polytechnic University, Saint Petersburg, Russia (e-mail: zelizaveta@yandex.ru).

$$\varphi = \varphi^* e^{-i\omega t}, \quad (4)$$

where ω is angular frequency of oscillation, t is a time, $\mathbf{u}^*, \varphi^*$ are complex amplitude.

Now we substitute equations (3) and (4) in the Lame's equations (1), (2). We get

$$(\mu+\alpha)\Delta\mathbf{u}^* - (\lambda+\mu-\alpha)grad\ div\mathbf{u}^* + \\ + 2\alpha rot\varphi^* + \omega^2\rho\mathbf{u}^* + \mathbf{X} = 0 \quad (5)$$

$$2\alpha rot\mathbf{u}^* + (\omega^2 J - 4\alpha)\varphi^* = 0. \quad (6)$$

In this paper we present a method for finding $\mathbf{u}^*, \varphi^*$, based on the use of the integral Fourier transform:

$$\hat{\mathbf{u}}^*(\mathbf{p}) = \int_V \mathbf{u}^*(\mathbf{R}) e^{-i\mathbf{p}\cdot\mathbf{R}} d\tau_R, \quad (7)$$

$$\hat{\varphi}^*(\mathbf{p}) = \int_V \varphi^*(\mathbf{R}) e^{-i\mathbf{p}\cdot\mathbf{R}} d\tau_R. \quad (8)$$

Here the integration is over the whole space V, the point of the source N_0 adopted as the origin of coordinates, **p**- variable vector of the Fourier transform. In this notation treatment theorem has the form:

$$\mathbf{u}^*(\mathbf{R}) = \frac{1}{(2\pi)^3}\int_V \hat{\mathbf{u}}^*(\mathbf{p}) e^{i\mathbf{p}\cdot\mathbf{R}} d\tau_p, \quad (9)$$

$$\varphi^*(\mathbf{R}) = \frac{1}{(2\pi)^3}\int_V \hat{\varphi}^*(\mathbf{p}) e^{i\mathbf{p}\cdot\mathbf{R}} d\tau_p. \quad (10)$$

Let us consider the case where the volume force X in the equation (5) is a unit concentrated pulsating force at the point of the source N_0. Now we can present the volume force as:

$\mathbf{X} = \mathbf{e}\delta(\mathbf{R})$, where

$$\delta(R) = \frac{1}{(2\pi)^3}\int_V e^{i\mathbf{p}\cdot\mathbf{R}} d\tau_p.$$

The Fourier transform of equations (5) and (6) have the following form:

$$(\rho\omega^2 - (\mu+\alpha)p^2)\hat{\mathbf{u}}^* - (\lambda+\mu-\alpha)\mathbf{pp}\cdot\hat{\mathbf{u}}^* + \\ + 2\alpha i\mathbf{p}\times\hat{\varphi}^* + \mathbf{e} = 0 \quad (11)$$

$$2\alpha i\mathbf{p}\times\hat{\mathbf{u}}^* + (J\omega^2 - 4\alpha)\hat{\varphi}^* = 0 \quad (12)$$

The translation vector $\mathbf{u}(N, N_0)$ and rotation vector $\varphi(N, N_0)$ of the observation point N are the fundamental solution of Lame's equations (5), (6). The point N is located in an unbounded space, which is loaded at the source point N_0 by the concentrated force **e** of the unit values.

So the solution can expressed as a following formula

$$\mathbf{u}^*(N, N_0) = \mathbf{U}(N, N_0)\cdot\mathbf{e}(N_0) \quad (13)$$

$$\varphi^*(N, N_0) = \mathbf{\Phi}(N, N_0)\cdot\mathbf{e}(N_0), \quad (14)$$

where $\mathbf{U}(N, N_0)$, $\mathbf{\Phi}(N, N_0)$ are second-rank tensors, (Kelvin-Somilyana's tensor).

Now the Fourier transform of equations (5) and (6) have the following form:

$$(\rho\omega^2 - (\mu+\alpha)p^2)\hat{\mathbf{U}} - (\lambda+\mu-\alpha)\mathbf{pp}\cdot\hat{\mathbf{U}} + \\ + 2\alpha i\mathbf{p}\times\hat{\mathbf{\Phi}} + \mathbf{E} = 0 \quad (15)$$

$$2\alpha i\mathbf{p}\times\hat{\mathbf{U}} + (J\omega^2 - 4\alpha)\hat{\mathbf{\Phi}} = 0. \quad (16)$$

We find $\mathbf{U}(N, N_0)$, $\mathbf{\Phi}(N, N_0)$ in the next form

$$\hat{\mathbf{U}}(N, N_0) = A\mathbf{E} + B\mathbf{pp} + C\mathbf{p}\times\mathbf{E}, \quad (17)$$

$$\hat{\mathbf{\Phi}}(N, N_0) = D\mathbf{E} + F\mathbf{pp} + G\mathbf{p}\times\mathbf{E}. \quad (18)$$

Let us substitute equations (17), (18) in (15), (16):

$$(\rho\omega^2 - (\mu+\alpha)p^2)A\mathbf{E} + (\rho\omega^2 - (\mu+\alpha)p^2)B\mathbf{pp} + \\ + (\rho\omega^2 - (\mu+\alpha)p^2)C\mathbf{p}\times\mathbf{E} - (\lambda+\mu-\alpha)A\mathbf{pp} - \\ - (\lambda+\mu-\alpha)Bp^2\mathbf{pp} + 2\alpha iD\mathbf{p}\times\mathbf{E} + 2\alpha iG\mathbf{p}\times\mathbf{E}\times p + \mathbf{E} = 0 \quad (19)$$

$$2\alpha iA\mathbf{p}\times\mathbf{E} + 2\alpha iC\mathbf{p}\times\mathbf{E}\times\mathbf{p} + (J\omega^2 - 4\alpha)D\mathbf{E} + \\ + (J\omega^2 - 4\alpha)F\mathbf{pp} + (J\omega^2 - 4\alpha)G\mathbf{p}\times\mathbf{E} = 0. \quad (20)$$

We get A,B,C,D,F,G because of independence **E,** $\mathbf{pp}$, $\mathbf{p}\times\mathbf{E}$.

$$A = \frac{1}{(\mu+\alpha)p^2 - \rho\omega^2},$$

$$B = -\frac{(\lambda+\mu-\alpha)A}{(\lambda+2\mu)p^2 - \rho\omega^2},$$

$$C = 0,$$

$$D = 0,$$

$$F = 0,$$

$$G = \frac{2i\alpha A}{4\alpha - J\omega^2}.$$

Substitution of these coefficients in the expression (17), (18), gives us

$$\hat{\mathbf{U}}(N, N_0) = \frac{1}{(\mu+\alpha)p^2 - \rho\omega^2}\mathbf{E} - \\ - \frac{(\lambda+\mu-\alpha)}{(\lambda+2\mu)p^2 - \rho\omega^2}\frac{1}{(\mu+\alpha)p^2 - \rho\omega^2}\mathbf{pp}, \quad (21)$$

$$\hat{\mathbf{\Phi}}(N, N_0) = \frac{2i\alpha}{4\alpha - J\omega^2}\frac{1}{(\mu+\alpha)p^2 - \rho\omega^2}\mathbf{p}\times\mathbf{E}. \quad (22)$$

Now we define

$$k_1^2 = \frac{\rho\omega^2}{\lambda+2\mu},$$

$$k_3^2 = \frac{\rho\omega^2}{\mu+\alpha},$$

$$\omega_0^2 = \frac{4\alpha}{J}.$$

And then we obtain

$$\hat{\mathbf{U}}(N,N_0) = \frac{1}{(\mu+\alpha)k_3^2}\left(\frac{k_3^2}{p^2-k_3^2}\mathbf{E} + \right.$$

$$\left. + \left(\frac{1}{p^2-k_1^2} - \frac{1}{p^2-k_3^2}\right)\mathbf{pp}\right) \quad (23)$$

$$\hat{\mathbf{\Phi}}(N,N_0) = \frac{i}{2(\mu+\alpha)}\frac{1}{1-\frac{\omega^2}{\omega_0^2}}\frac{1}{p^2-k_3^2}\mathbf{p}\times\mathbf{E} \quad (24)$$

Using (9) and (10) we arrive at

$$\mathbf{U} = \frac{1}{4\pi(\mu+\alpha)k_3^2}(k_3^2\mathbf{E}f_3 + \nabla\nabla(f_1 - f_3)), \quad (25)$$

$$\Phi = \frac{1}{8(\mu+\alpha)}\frac{1}{1-\frac{\omega^2}{\omega_0^2}}\nabla f_3\times\mathbf{E}, \quad (26)$$

where

$$f_s(\mathbf{R}) = \frac{1}{2\pi^2}\int_V \frac{e^{i\mathbf{p}\cdot\mathbf{R}}}{p^2-k_s^2}d\tau_p. \quad (27)$$

f_s satisfies the inhomogeneous Helmholtz equation:

$$\Delta f_s + k_s^2 f_s = -4\pi\delta(\mathbf{R}). \quad (28)$$

So, f_s is the fundamental solution χ_s determined by the formula

$$\chi_s = \frac{e^{ik_sR}}{R}. \quad (29)$$

Let us substitute $f_s = \chi_s$ in (25) and (26), we obtain the following result

$$\mathbf{U} = \frac{1}{4\pi(\mu+\alpha)k_3^2}\left(k_3^2\mathbf{E}\frac{e^{ik_3R}}{R} + \nabla\nabla\left(\frac{e^{ik_1R}}{R} - \frac{e^{ik_3R}}{R}\right)\right) \quad (30)$$

$$\Phi = \frac{1}{8(\mu+\alpha)}\frac{1}{1-\frac{\omega^2}{\omega_0^2}}\nabla\frac{e^{ik_3R}}{R}\times\mathbf{E} \quad (31)$$

Thus, taking into account (13), (14), (3) and (4) we find the displacement vector and rotation vector in a reduced Cosserat continuum from the action harmonic unit force. Harmonic force in our study is a unit concentrated pulsating force.

III. Conclusion

In this work we consider reduced Cosserat continuum as a model for granular medium. In this continuum translations and rotations are independent, stress tensor is not symmetric and couple stresses tensor equal to zero.

We found the fundamental solution: translation and rotation vectors caused by arbitrarily directed concentrated force, harmonically dependent on time and having a unit amplitude at a given point of a homogeneous, unbounded, isotropic elastic medium.

$$\mathbf{u} = \mathbf{U}(N,N_0)\cdot\mathbf{e}(N_0)e^{-i\omega t},$$

$$\varphi = \mathbf{\Phi}(N,N_0)\cdot\mathbf{e}(N_0)e^{-i\omega t},$$

where $\mathbf{U}(N,N_0)$, $\mathbf{\Phi}(N,N_0)$ is presented in equation (30), (31).

The problem is solved with the use of Fourier transforms. In many problems of the body force acting at a point - an adequate model of the seismic source displacement observed during the earthquake.

References

[1] S. Skuodis, A. Norkus, L. Tumonis, J. Amsiejus, C. Aksamitauskas "Experimental and numerical investigation of sand compression pecullarities" in *Journal of civil engineering and management.* Vol 19(1), 2013. Pp78-85.

[2] E.N. Kurbatskiy, O.A. Golosova " Features of the propagation of stress waves in natural and artificial granular media" in *Structural mechanics and calculation of structures.* №2, 2011. pp. 45-50. (in Russian)

[3] Badanin A., Bugrov A., Krotov A. The determination of the first critical load on particulate medium of sandy loam foundation // Magazine of Civil Engineering. 9. 2012. Pp 29-34. (in Russian).

[4] Heinrich M. Jaeger, Sidney R. Nagel. "Granular solids, liquids, and gases" *Reviews of modern physics.* vol.68, №4, 1996. pp 1259-1273.

[5] Arslan H., Sture S. "Finite element simulation of localization in granular materials by micropolar continuum approach" in *Computers and Geotechnics.* 35 (4), 2008. pp. 548-562.

[6] Harris D. "Double-slip and Spin: A generalisation of the plastic potential model in the mechanics of granular materials" in *International Journal of Engineering Science*. Volume 47, Issues 11–12, 2009. pp.208-1215.

[7] Prosvetov, V.I., Sumets, P.P., Verveyko, N.D. "Modeling of flow of medium with homogeneous microstructure" in *International Journal of Mathematical Models and Methods in Applied Sciences*. V.5 (3),2011. pp. 508-516

[8] Neff P., Jeong J. "A new paradigm: the linear isotropic Cosserat model with conformally invariant curvature energy" in *ZAMM Zeitschrift fur Angewandte Mathematik und Mechanik.* vol.89, №2, February 2009, pp 107-122.

[9] Jasiuk I., Ostoja-Starzewski M. "On the reduction of constants in planar cosserat elasticity with eigenstrains and eigencurvatures" in *Journal of Thermal Stresses*. 26 (11-12), 2003. pp.1221-1228.

[10] Jeong J., Neff P. "Existence, uniqueness and stability in linear cosserat elasticity for weakest curvature conditions" in *Mathematics and Mechanics of Solids.* 15 (1), 2010. pp. 78-95.

[11] Schwartz L.M., Johnson D.L., Feng S. "Vibrational modes in granular materials" in *Physical review letters*, v. 52, №10, 1984. pp.831-834.

[12] Grekova E.F., Kulesh M.A., Herman G.C. "Waves in linear elastic media with microrotations, part 2: Isotropic reduced Cosserat model" in *Bulletin of the Seismological Society of America*. 99 (2 B), 2009. pp. 1423-1428.

[13] Grekova E.F."Nonlinear isotropic elastic reduced Cosserat continuum as a possible model for geomedium and geomaterials. Spherical prestressed state in the semilinear material" in *Journal of seismology*. . vol 16, issue 4, 2012. pp695-707.

[14] Lalin V., Zdanchuk E. "Reduced Cosserat continuum as a possible model for granulsr medium" in *Proceedings of the International Conference „Innovative Materials, Structures and Technologies", Riga. 2014.*

[15] Lalin V., Zdanchuk E. “Nonlinear thermodynamic model for granular medium” *Recent Advances in Mechanical Engineering and Mechanics.Proceedings of the 2014 International Conference on Theoretical Mechanics and Applied Mechanics* , Venice, Italy. 2014 Pp32-35.

[16] Lalin V., Zdanchuk E. “Nonlinear thermodynamic model for reduced Cosserat continuum” in *International Journal of Mathematical Models and Methods in Applied Sciences,* V. 8. 2014. Pp 208-213

[17] Keiiti Aki, Paul G. Richards *Quantitative seismology. Theory and methods*. W.H. Freeman and company. San Francisco. 1980.

The 2-Point Explicit Group Successive Over-Relaxation Method for Solving Fredholm Integral Equations of the Second Kind

Mohana Sundaram Muthuvalu, Elayaraja Aruchunan, Jumat Sulaiman, Samsul Ariffin Abdul Karim and Mohammad Mehdi Rashidi

***Abstract*—**In this paper, performance analysis one of the Explicit Group (EG) methods i.e. 2-Point Explicit Group Successive Over-Relaxation (2-EGSOR) will be investigated. The 2-EGSOR method will be applied sequentially in solving first order composite closed Newton-Cotes (1-CCNC) algebraic equations associated with the numerical solutions of the linear Fredholm integral equations of the second kind. Numerical results are included in order to verify the performance of the method compared with the 2-Point Explicit Group Gauss-Seidel (2-EGGS) method. Based on the numerical results obtained, the results show that 2-EGSOR method is better than 2-EGGS method in terms of number of iterations and CPU time.

***Keywords*—**Fredholm integral equations, Explicit Group methods, Composite closed Newton-Cotes scheme, Dense linear system.

I. Introduction

INTEGRAL equations (IEs) have been one of the principal mathematical models in various areas of science and engineering. The IEs are encountered in numerous applications including continuum mechanics, potential theory, geophysics, electricity and magnetism, kinetic theory of gases, hereditary phenomena in physics and biology, renewal theory, quantum mechanics, radiation, optimization, optimal control systems, communication theory, mathematical economics, population genetics, queuing theory, medicine, mathematical problems of radiative equilibrium, particle transport problems of astrophysics and reactor theory, acoustics, fluid mechanics, steady state heat conduction, fracture mechanics and radiative heat transfer problems [1]. Consequently, in this paper, a type of IEs i.e. linear Fredholm integral equations of the second kind is considered.

The general form of linear Fredholm integral equations of the second kind can be defined as follows

$$\varphi(x) - \int_{\alpha}^{\beta} K(x,t)\varphi(t)dt = f(x), \; x \in [\alpha, \beta]. \qquad (1)$$

The right-hand side function f and kernel K are given. Meanwhile, φ is the unknown function to be determined. The kernel function K is assumed to be absolutely integrable and satisfy the Fredholm alternative theorem [2]. The application of numerical methods for solving the problem (1) is the focus of this paper. There is a huge literature on numerical methods for solving problem (1), for instance refer [3]-[7]. The implementations of numerical methods on problem (1) mostly lead to dense linear systems. Thus, efficient iterative solvers are required to solve the resulting dense linear systems.

Recently, a family of block iterative methods known as Explicit Group (EG) iterative methods has been applied widely in solving various types of linear systems. Thus, in this paper, performance of an iterative method under EG methods i.e. 2-Point Explicit Group Successive Over-Relaxation (2-EGSOR) will be investigated in solving first order composite closed Newton-Cotes quadrature (1-CCNC) algebraic equations. The performance of the 2-EGSOR method on 1-CCNC algebraic equations is comparatively studied by their application in solving problem (1). The concept of the 2-EGSOR method is derived by combining the standard 2-Point Explicit Group (2-EG) method with Successive Over-Relaxation (SOR) approach. Numerical performance of the 2-EGSOR method will be compared with the standard 2-EG method. The standard 2-EG method is also known as 2-Point Explicit Group Gauss-Seidel (2-EGGS) method.

This paper is organised in five main sections. Section II explains the derivation of 1-CCNC algebraic equations for problem (1) followed by the formulations of the 2-EGGS and

M. S. Muthuvalu is with the Department of Petroleum Engineering, Faculty of Geosciences and Petroleum Engineering, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia (corresponding author; phone: +605-3687695; e-mail: msmuthuvalu@gmail.com).

E. Aruchunan is with the Department of Mathematics and Statistics, Curtin University, Perth WA6845, Australia (e-mail: earuchunan@yahoo.com).

J. Sulaiman is with the School of Science and Technology, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia (e-mail: jumat@ums.edu.my).

S. A. A. Karim is with the Department of Fundamental and Applied Sciences, Faculty of Science and Information Technology, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia (e-mail: samsul_ariffin@petronas.com.my).

M. M. Rashidi is with the Mechanical Engineering Department, University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai, Peoples Republic of China and Mechanical Engineering Department, Engineering Faculty of Bu-Ali Sina University, Hamedan, Iran (e-mail: mm_rashidi@sjtu.edu.cn).

2-EGSOR methods in Section III. In Section IV, results based on numerical simulations are presented to assert the performance of the tested methods. Finally, the concluding remarks are given in Section V.

II. 1-CCNC Algebraic Equations

In this section, discretisation of the problem (1) by using 1-CCNC scheme is discussed. An application of the 1-CCNC scheme for problem (1) leads to 1-CCNC algebraic equations which will be solved by using 2-EGGS and 2-EGSOR methods. Now, let the interval $[\alpha,\beta]$ divided uniformly into N subintervals and the discrete set of points of x and t given by $x_i = \alpha + ih \quad (i = 0,1,2,\cdots,N-2,N-1,N)$ and $t_j = \alpha + jh \quad (j = 0,1,2,\cdots,N-2,N-1,N)$ respectively, where the constant step size, h is defined as follows

$$h = \frac{\beta - \alpha}{N}. \tag{2}$$

Before further explanations, the following notations i.e., $K_{i,j} \equiv K(x_i,t_j)$, $\hat{\varphi}_i \equiv \hat{\varphi}(x_i)$, $\hat{\varphi}_j \equiv \hat{\varphi}(t_j)$ and $f_i \equiv f(x_i)$ will be applied subsequently for simplicity.

An application of the 1-CCNC scheme reduces problem (1) into algebraic equations as follows [5], [8]

$$\hat{\varphi}_i - \sum_{j=0}^{N} w_j K_{i,j} \hat{\varphi}_j = f_i \tag{3}$$

for $i = 0,1,2,\cdots,N-2,N-1,N$. The solution $\hat{\varphi}$ is an approximation of the exact solution φ to (1) and w_j is the weights of 1-CCNC scheme that satisfies the following condition

$$w_j = \begin{cases} \dfrac{h}{2}, & j = 0,N \\ h, & otherwise \end{cases}. \tag{4}$$

Following the conventional process, the generated 1-CCNC algebraic equations (3) can be written as the following matrix form

$$A\hat{\varphi} = f. \tag{5}$$

where $A = [a_{i,j}] \in \Re^{(N+1)\times(N+1)}$ is a real and dense coefficient matrix with elements

$$a_{i,j} = \begin{cases} 1 - w_j K_{i,j}, & i = j \\ -w_j K_{i,j}, & i \neq j \end{cases}. \tag{6}$$

III. 2-EGGS and 2-EGSOR Iterative Methods

As afore-mentioned, the formulation and implementation of the 2-EGGS and 2-EGSOR methods for solving the generated 1-CCNC algebraic equations will be discussed. Now, let consider any group of two points i.e., x_i and x_{i+1} that are used simultaneously to calculate the values of $\hat{\varphi}$ based on algebraic equations (3). Therefore, at point x_i, the solution is approximated by

$$\hat{\varphi}_i - \sum_{j=0}^{N} w_j K_{i,j} \hat{\varphi}_j = f_i \text{ (i.e. equation (3))}. \tag{7}$$

Whereas, at point x_{i+1} the solution is given by

$$\hat{\varphi}_{i+1} - \sum_{j=0}^{N} w_j K_{i+1,j} \hat{\varphi}_j = f_{i+1}. \tag{8}$$

Now, the equations (7) and (8) can be written simultaneously in the matrix form as follows

$$\begin{bmatrix} a_{i,i} & a_{i,i+1} \\ a_{i+1,i} & a_{i+1,i+1} \end{bmatrix} \begin{bmatrix} \hat{\varphi}_i \\ \hat{\varphi}_{i+1} \end{bmatrix} = \begin{bmatrix} f_i - \sum_{j=0}^{i-1} a_{i,j} \hat{\varphi}_j - \sum_{j=i+2}^{N} a_{i,j} \hat{\varphi}_j \\ f_{i+1} - \sum_{j=0}^{i-1} a_{i+1,j} \hat{\varphi}_j - \sum_{j=i+2}^{N} a_{i+1,j} \hat{\varphi}_j \end{bmatrix} \tag{9}$$

where coefficient matrix with size (2×2) can be easily inverted. Thus, the equation (9) can be written in explicit form as

$$\begin{bmatrix} \hat{\varphi}_i \\ \hat{\varphi}_{i+1} \end{bmatrix} = \frac{1}{|B|} \begin{bmatrix} a_{i+1,i+1} & -a_{i,i+1} \\ -a_{i+1,i} & a_{i,i} \end{bmatrix} \begin{bmatrix} f_i - \sum_{j=0}^{i-1} a_{i,j} \hat{\varphi}_j - \sum_{j=i+2}^{N} a_{i,j} \hat{\varphi}_j \\ f_{i+1} - \sum_{j=0}^{i-1} a_{i+1,j} \hat{\varphi}_j - \sum_{j=i+2}^{N} a_{i+1,j} \hat{\varphi}_j \end{bmatrix} \tag{10}$$

where $|B| = \det B = a_{i,i}a_{i+1,i+1} - a_{i+1,i}a_{i,i+1}$. This simplifies to the formulae

$$\begin{bmatrix} \hat{\varphi}_i \\ \hat{\varphi}_{i+1} \end{bmatrix} = \frac{1}{|B|} \begin{bmatrix} a_{i+1,i+1}(C) - a_{i,i+1}(D) \\ -a_{i+1,i}(C) + a_{i,i}(D) \end{bmatrix} \tag{11}$$

with

$$C = f_i - \sum_{j=0}^{i-1} a_{i,j} \hat{\varphi}_j - \sum_{j=i+2}^{N} a_{i,j} \hat{\varphi}_j \tag{12}$$

and

$$D = f_{i+1} - \sum_{j=0}^{i-1} a_{i+1,j} \hat{\varphi}_j - \sum_{j=i+2}^{N} a_{i+1,j} \hat{\varphi}_j . \tag{13}$$

Hence, the iterative scheme for 2-EGGS method is given by

$$\begin{bmatrix} \hat{\varphi}_i \\ \hat{\varphi}_{i+1} \end{bmatrix}^{(k+1)} = \frac{1}{|B|} \begin{bmatrix} a_{i+1,i+1}(C) - a_{i,i+1}(D) \\ -a_{i+1,i}(C) + a_{i,i}(D) \end{bmatrix} \tag{14}$$

for $i = 0,2,4,\cdots,N-3,N-1$, where

$$C = f_i - \sum_{j=0}^{i-1} a_{i,j} \hat{\varphi}_j^{(k+1)} - \sum_{j=i+2}^{N} a_{i,j} \hat{\varphi}_j^{(k)} \tag{15}$$

and

$$D = f_{i+1} - \sum_{j=0}^{i-1} a_{i+1,j} \hat{\varphi}_j^{(k+1)} - \sum_{j=i+2}^{N} a_{i+1,j} \hat{\varphi}_j^{(k)} . \tag{16}$$

By adding an accelerated parameter, ω into formulae (11), the iterative scheme for 2-EGSOR method can be rewritten as

$$\begin{bmatrix} \hat{\varphi}_i \\ \hat{\varphi}_{i+1} \end{bmatrix}^{(k+1)} = (1-\omega) \begin{bmatrix} \hat{\varphi}_i \\ \hat{\varphi}_{i+1} \end{bmatrix}^{(k)} + \frac{\omega}{|B|} \begin{bmatrix} a_{i+1,i+1}(C) - a_{i,i+1}(D) \\ -a_{i+1,i}(C) + a_{i,i}(D) \end{bmatrix} \tag{17}$$

for $i = 0,2,4,\cdots,N-3,N-1$, where C and D are as shown in equations (15) and (16) respectively.

For an even subintervals, N, the number of discrete node points is odd i.e., $N+1$, which results in one ungrouped point. Therefore, the ungrouped point i.e., x_N, will be computed based on the following point iterations

$$\hat{\varphi}_N^{(k+1)} = \frac{1}{a_{N,N}} \left[f_i - \sum_{j=0}^{N-1} \left(a_{N,j} \hat{\varphi}_j^{(k+1)} \right) \right] \tag{18}$$

and

$$\hat{\varphi}_N^{(k+1)} = (1-\omega) \hat{\varphi}_N^{(k)} + \frac{\omega}{a_{N,N}} \left[f_i - \sum_{j=0}^{N-1} \left(a_{N,j} \hat{\varphi}_j^{(k+1)} \right) \right] \tag{19}$$

for 2-EGGS and 2-EGSOR methods respectively. For the convergence test, an absolute criterion depending on the iteration error i.e., $\left\| \hat{\varphi}^{(k+1)} - \hat{\varphi}^{(k)} \right\| \le \varepsilon$ (where ε is the convergence criterion) is applied for both methods.

IV. Simulation Results

The following two linear Fredholm integral equations of the second kind are used as the test problems in order to compare the performance of the methods.

Test Problem 1 [1]
Consider the Fredholm integral equation of the second kind

$$\varphi(x) - \int_0^1 \left(4xt - x^2\right) \varphi(t) dt = x , \; x \in [0,1], \tag{20}$$

and the exact solution is given by

$$\varphi(x) = 24x - 9x^2 .$$

Test Problem 2 [5]
Consider the Fredholm integral equation of the second kind

$$\varphi(x) - \int_0^1 \left(x^2 + t^2\right) \varphi(t) dt = x^6 - 5x^3 + x + 10 , \; x \in [0,1], \tag{21}$$

with the exact solution

$$\varphi(x) = x^6 - 5x^3 + \frac{1045}{28} x^2 + x + \frac{2141}{84} .$$

For the numerical simulations, three criteria are used for a comparative analysis i.e.
k - Number of iterations,
CPU - CPU time (in seconds) when the converged solution is obtained,
$RMSE$ - Root mean square error [9].

The value of initial datum, $\hat{\varphi}^{(0)}$, is set to zero for both the test problems. The computations are performed on a personal computer with Intel(R) Core(TM) i3-2120 CPU and 4.00GB RAM, and the programming codes are compiled by using C language. Throughout the simulations, the convergence test considered is $\varepsilon = 10^{-10}$ and tested on eight different values of N i.e. 60, 120, 240, 480, 960, 1920, 3840 and 7680. Meanwhile, the experimental values of ω were obtained within ± 0.01 by running the programs for different values of

ω and choosing the one that gave the minimum number of iterations. For the case of more than one ω (based on minimum number of iterations), the optimum value of ω is chosen by considering the minimum *RMSE*. The numerical results of the tested methods for test problems 1 and 2 are presented in Tables I and II respectively.

TABLE I. NUMERICAL RESULTS OF TEST PROBLEM 1

N	*Methods*	*k*	*CPU*	*RMSE*
	2-EGGS	183	0.13	2.29894×10^{-02}
60	2-EGSOR	40	0.04	2.29894×10^{-02}
		($\omega = 1.53$)		
	2-EGGS	189	0.31	5.71079×10^{-03}
120	2-EGSOR	40	0.06	5.71079×10^{-03}
		($\omega = 1.54$)		
	2-EGGS	192	1.10	1.42375×10^{-03}
240	2-EGSOR	40	0.25	1.42375×10^{-03}
		($\omega = 1.54$)		
	2-EGGS	193	4.28	3.55480×10^{-04}
480	2-EGSOR	41	0.94	3.55481×10^{-04}
		($\omega = 1.54$)		
	2-EGGS	194	17.02	8.88150×10^{-05}
960	2-EGSOR	41	3.69	8.88154×10^{-05}
		($\omega = 1.54$)		
	2-EGGS	194	67.92	2.21967×10^{-05}
1920	2-EGSOR	41	14.64	2.21972×10^{-05}
		($\omega = 1.54$)		
	2-EGGS	195	274.62	5.54797×10^{-06}
3840	2-EGSOR	41	58.48	5.54847×10^{-06}
		($\omega = 1.55$)		
	2-EGGS	195	1091.82	1.38651×10^{-06}
7680	2-EGSOR	41	235.13	1.38702×10^{-06}
		($\omega = 1.55$)		

TABLE II. NUMERICAL RESULTS OF TEST PROBLEM 2

N	*Methods*	*k*	*CPU*	*RMSE*
	2-EGGS	54	0.05	2.15612×10^{-02}
60	2-EGSOR	23	0.02	2.15612×10^{-02}
		($\omega = 1.27$)		
	2-EGGS	55	0.09	5.35413×10^{-03}
120	2-EGSOR	23	0.05	5.35413×10^{-03}
		($\omega = 1.28$)		
	2-EGGS	55	0.32	1.33417×10^{-03}
240	2-EGSOR	23	0.14	1.33417×10^{-03}
		($\omega = 1.28$)		
	2-EGGS	56	1.24	3.33005×10^{-04}
480	2-EGSOR	23	0.56	3.33005×10^{-04}
		($\omega = 1.28$)		
	2-EGGS	56	4.92	8.31845×10^{-05}
960	2-EGSOR	23	2.06	8.31846×10^{-05}
		($\omega = 1.29$)		
	2-EGGS	56	19.81	2.07878×10^{-05}
1920	2-EGSOR	23	8.06	2.07879×10^{-05}
		($\omega = 1.29$)		
	2-EGGS	56	79.49	5.19584×10^{-06}
3840	2-EGSOR	23	31.89	5.19593×10^{-06}
		($\omega = 1.29$)		
	2-EGGS	56	317.22	1.29876×10^{-06}
7680	2-EGSOR	23	127.17	1.29885×10^{-06}
		($\omega = 1.29$)		

V. CONCLUSION

In this paper, 2-EGSOR method has been successfully applied in solving linear Fredholm integral equations of the second kind. By referring Tables I and II, the numerical results show that implementation of the 2-EGSOR method solved the both test problems with minimum number of iterations and fastest CPU time. In terms of accuracy, numerical solutions obtained via 2-EGSOR method are in good agreement compared to the 2-EGGS method. Finally, it can be summarized that the 2-EGSOR method is better than 2-EGGS method, especially in the aspect of number of iterations and CPU time.

ACKNOWLEDGMENT

The first author gratefully acknowledges the financial support received from the Universiti Teknologi PETRONAS for this research work.

REFERENCES

[1] W. Wang, "A new mechanical algorithm for solving the second kind of Fredholm integral equation," *Appl. Math. Comput.*, vol. 172, pp. 946–962, 2006.

[2] K. E. Atkinson, *The numerical solution of integral equations of the second kind*. Cambridge University Press, 1997.

[3] A. Chakrabarti, and S. C. Martha, "Approximate solutions of Fredholm integral equations of the second kind," *Appl. Math. Comput.*, vol. 211, pp. 459–466, 2009.

[4] G. Mastroianni and G. V. Milovanović, "Well-conditioned matrices for numerical treatment of Fredholm integral equations of the second kind," *Numer. Linear Algebra Appl.*, vol. 16, pp. 995–1011, 2009.

[5] M. S. Muthuvalu and J. Sulaiman, "Half-Sweep Arithmetic Mean method with composite trapezoidal scheme for solving linear Fredholm integral equations," *Appl. Math. Comput.*, vol. 217, pp. 5442–5448, 2011.

[6] J. Rashidinia and Z. Mahmoodi, "Collocation method for Fredholm and Volterra integral equations," *Kybernetes*, vol. 42, pp. 400–412, 2013.

[7] X. –C. Zhong, "A new Nyström-type method for Fredholm integral equations of the second kind," *Appl. Math. Comput.*, vol. 219, pp. 8842–8847, 2013.

[8] M. S. Muthuvalu and J. Sulaiman, "The Quarter-Sweep Geometric Mean method for solving second kind linear Fredholm integral equations," *Bull. Malays. Math. Sci. Soc.*, vol. 36, pp. 1009–1026, 2013.

[9] A. Golbabai and S. Seifollahi, "An iterative solution for the second kind integral equations using radial basis functions," *Appl. Math. Comput.*, vol. 181, pp. 903–907, 2006.

Integrated mathematical model of the engine and the aircraft longitudinal dynamics

Constantin Rotaru, and Ionică Cîrciu

***Abstract*—**This paper presents a mathematical model of the integrated engine and the aircraft longitudinal dynamics based on the theory of linear and nonlinear systems. The dynamics of the engine was represented by a linear, time variant model near a nominal operating point within a finite time interval. The linearized equations were expressed in a matrix form, suitable for the incorporation in the MAPLE program solver. The behavior of the engine was included in terms of variation of the rotational speed following a deflection of the throttle. The aircraft and engine models were incorporated into a single state variable model.

***Keywords*—**aircraft, engine control, Laplace transform, turbojet.

I. Introduction

THE aircraft engine must provide a wide range of predictable and repeatable thrust performance over the entire operating envelope of the engine, which can cover the altitude from sea level to tens of thousands meters. These altitude changes along with variations in flight speed from takeoff to supersonic velocities result in large, simultaneous variations in engine inlet temperature, inlet pressure and exhaust pressure. These large variations in engine operating conditions and the demand for precise thrust control, coupled with the demand for highly reliable operations, despite the complexity of the engine itself, create a significant challenge for the design of the engine control systems.

Modern gas turbine engine control systems are closed-loop control systems that consist of all four types of control components: controller, sensor, actuator and accessory. The simplest engine control system is one that produces desired engine thrust or shaft power by changing the fuel flow. Because reliable, in-flight engine thrust measurement is not currently practical, the engine shaft rotational speed (N) or engine pressure ratio (EPR) has been used effectively as an indicator of engine thrust (or power). Hence, for this simplified control system, the command variable (or the desired output variable) is shaft speed (or engine pressure ratio), the control variable is actuator position, the actuator is fuel metering valve, the output of the metering valve is the fuel flow that is injected in the combustor, the output of the engine is engine power setting variable (shaft speed or engine pressure ratio); furthermore, fuel control accessory components are the fuel tank, the fuel pump and the sensors (tachometer or pressure transducers).

In general terms, a generic *n*-th order dynamic system with multiple inputs and instrumented with sensors providing measurements for the output variables can be described by the following system of differential equations:

$$\begin{cases} \dot{x}_1(t) = f_1(x_1(t), x_2(t), ...x_n(t), u_1(t), u_2(t)....u_m(t)) \\ \dot{x}_2(t) = f_2(x_1(t), x_2(t), ...x_n(t), u_1(t), u_2(t)....u_m(t)) \\ ..... \\ \dot{x}_2(t) = f_n(x_1(t), x_2(t), ...x_n(t), u_1(t), u_2(t)....u_m(t)) \end{cases} \quad (1)$$

$$\begin{cases} y_1(t) = g_1(x_1(t), x_2(t), ...x_n(t), u_1(t), u_2(t)....u_m(t)) \\ y_2(t) = g_2(x_1(t), x_2(t), ...x_n(t), u_1(t), u_2(t)....u_m(t)) \\ ..... \\ y_l(t) = g_l(x_1(t), x_2(t), ...x_n(t), u_1(t), u_2(t)....u_m(t)) \end{cases} \quad (2)$$

where $x_1(t), x_2(t), x_3(t), ...x_n(t)$ are the state variables, $u_1(t), u_2(t), u_3(t), ...u_m(t)$ are the system inputs and $y_1(t), y_2(t), y_3(t), ...y_l(t)$ are the system outputs. For nonlinear dynamic systems with multiple inputs and multiple outputs, the functions $f_1, f_2, ...f_n$ and $g_1, g_2, ...g_n$ are nonlinear. If the system can be linearized around a set of operating initial conditions, the state variable model reduces itself to the form [1]:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ .... \\ \dot{x}_n(t) \end{bmatrix} = A_{n\times n} \begin{bmatrix} x_1(t) \\ x_2(t) \\ .... \\ x_n(t) \end{bmatrix} + B_{n\times m} \begin{bmatrix} u_1(t) \\ u_2(t) \\ .... \\ u_m(t) \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ .... \\ y_l(t) \end{bmatrix} = C_{l\times n} \begin{bmatrix} x_1(t) \\ x_2(t) \\ .... \\ x_n(t) \end{bmatrix} + D_{l\times m} \begin{bmatrix} u_1(t) \\ u_2(t) \\ .... \\ u_m(t) \end{bmatrix} \quad (4)$$

Constantin Rotaru is with the Aviation Integrated Systems and Mechanics Department, Military Technical Academy, Bucharest, 050141, Romania (corresponding author, tel: +40745974488; fax: +4021 335 57 63; e-mail: rotaruconstantin@yahoo.com).

Ionică Cîrciu is with the Department of Aviation, "Henri Coandă" Air Force Academy, Braşov 500187, Romania (e-mail: circiuionica@yahoo.co.uk).

Using the superimposition of effects associated with the linearity of the system, the Laplace transform of the *i*-th output of the system is given by

$$\begin{aligned} Y_i(s) &= \left(\frac{Y_i(s)}{U_1(s)}\right)\cdot U_1(s) + \left(\frac{Y_i(s)}{U_2(s)}\right)\cdot U_2(s) + ... \left(\frac{Y_i(s)}{U_m(s)}\right)\cdot U_m(s) \\ &= G_{i1}(s)\cdot U_1(s) + G_{i2}(s)\cdot U_2(s) + ...G_{im}(s)\cdot U_m(s) \end{aligned} \quad (5)$$

Considering all the measureable outputs subjected to inputs we have

$$\begin{cases} Y_1(s) = G_{11}(s)\cdot U_1(s) + G_{12}(s)\cdot U_2(s) + ...G_{1m}(s)\cdot U_m(s) \\ Y_2(s) = G_{21}(s)\cdot U_1(s) + G_{22}(s)\cdot U_2(s) + ...G_{2m}(s)\cdot U_m(s) \\ ..... \\ Y_i(s) = G_{i1}(s)\cdot U_1(s) + G_{i2}(s)\cdot U_2(s) + ...G_{im}(s)\cdot U_m(s) \\ ..... \\ Y_l(s) = G_{l1}(s)\cdot U_1(s) + G_{l2}(s)\cdot U_2(s) + ...G_{lm}(s)\cdot U_m(s) \end{cases} \quad (6)$$

With the following notation

$$\left[\overline{Y}(s)\right]_{l\times 1} = \begin{bmatrix} Y_1(s) \\ Y_2(s) \\ ..... \\ Y_i(s) \\ ..... \\ Y_l(s) \end{bmatrix}_{l\times 1} \quad \text{and} \quad \left[U(s)\right]_{m\times 1} = \begin{bmatrix} U_1(s) \\ U_2(s) \\ ..... \\ U_i(s) \\ ..... \\ U_m(s) \end{bmatrix}_{m\times 1} \quad (7)$$

the matrix of transfer function can takes on the form

$$\frac{\left[Y(s)\right]_{l\times 1}}{\left[U(s)\right]_{m\times 1}} = \begin{bmatrix} G_{11}(s) & G_{12}(s) & ... & G_{1i}(s) & ... & G_{11}(s) \\ G_{11}(s) & G_{11}(s) & ... & G_{11}(s) & ... & G_{11}(s) \\ ... & ... & ... & ... & ... & ... \\ G_{11}(s) & G_{11}(s) & ... & G_{11}(s) & ... & G_{11}(s) \\ ... & ... & ... & ... & ... & ... \\ G_{11}(s) & G_{11}(s) & ... & G_{11}(s) & ... & G_{11}(s) \end{bmatrix} \quad (8)$$

Based on this modeling, it is possible to derive a closed-form relationship between the state variable model of the system and its matrix transfer functions. The application of the Laplace transformation to the state variable model, with zero initial conditions

$$\begin{cases} \dot{x} = A\cdot x + B\cdot u \\ y = C\cdot x + D\cdot u \end{cases} \quad (9)$$

leads to

$$\begin{cases} sX(s) = A\cdot X(s) + B\cdot U(s) \\ Y(s) = C\cdot X(s) + D\cdot U(s) \end{cases} \quad (10)$$

Solving for $X(s)$ from the first relationship above, we have

$$Y(s) = \left[C(sI - A)^{-1}B + D\right]U(s) \quad (11)$$

where *I* is the identity matrix. The transfer function of the state variable model is

$$\frac{Y(s)}{U(s)} = G(s) = \left[C(sI - A)^{-1}B + D\right] \quad (12)$$

The roots of the characteristic equation of the system are also the eigenvalues of the state matrix *A*.

II. STATE VARIABLE MODELING OF THE AIRCRAFT LONGITUDINAL DYNAMICS

The starting point for the development of the state variable model of the aircraft longitudinal dynamics is the set of longitudinal equations

$$\begin{cases} \dot{u} = (X_u + X_{Tu})u + X_\alpha\alpha - g\theta\cos(\Theta_1) + X_{\delta_E}\delta_E \\ U_1\dot{\alpha} = Z_u u + Z_\alpha\alpha + Z_{\dot{\alpha}}\dot{\alpha} - g\theta\sin(\Theta_1) + (Z_q + U_1)\dot{\theta} + Z_{\delta_E}\delta_E \\ \ddot{\theta} = (M_u + M_{Tu})u + (M_\alpha + M_{T\alpha})\alpha + M_{\dot{\alpha}}\dot{\alpha} + M_q\dot{\theta} + M_{\delta_E}\delta_E \end{cases} \quad (13)$$

where

$$X_u, X_{Tu}, X_\alpha, X_{\delta_E}, Z_u, Z_\alpha, Z_{\dot{\alpha}}, Z_q,$$
$$Z_{\delta_E}, M_u, M_{Tu}, M_\alpha, M_{\delta_E}, M_{T\alpha}, M_q$$

are the longitudinal dimensional stability and control derivatives and $u, \alpha, \theta, \Theta_1, \delta_E$ are the perturbation velocity, angle of attack, pitch and elevator angles.

Using the relationship $q = \dot{\theta}$ şi $\dot{q} = \ddot{\theta}$, the above equations can be rewritten as

$$\begin{cases} \dot{u} = (X_u + X_{Tu})u + X_\alpha\alpha - g\theta\cos(\Theta_1) + X_{\delta_E}\delta_E \\ (U_1 - Z_{\dot{\alpha}})\dot{\alpha} = Z_u u + Z_\alpha\alpha - g\theta\sin(\Theta_1) + (Z_q + U_1)q + Z_{\delta_E}\delta_E \\ \dot{q} = (M_u + M_{Tu})u + (M_\alpha + M_{T\alpha})\alpha + M_{\dot{\alpha}}\dot{\alpha} + M_q q + M_{\delta_E}\delta_E \\ \dot{\theta} = q \end{cases} \quad (14)$$

The second equation in $\dot{\alpha}$ is embedded into the third equation in *q*. Solving the second equation and inserting it in the third equations, leads to

$$\begin{cases} \dot{u} = (X_u + X_{Tu})u + X_\alpha\alpha - g\theta\cos(\Theta_1) + X_{\delta_E}\delta_E \\ \dot{\alpha} = \dfrac{Z_u}{(U_1 - Z_{\dot{\alpha}})}u + \dfrac{Z_\alpha}{(U_1 - Z_{\dot{\alpha}})}\alpha - \dfrac{g\sin(\Theta_1)}{(U_1 - Z_{\dot{\alpha}})}\theta + \\ \quad + \dfrac{(Z_q + U_1)}{(U_1 - Z_{\dot{\alpha}})}\dot{\theta} + \dfrac{Z_{\delta_E}}{(U_1 - Z_{\dot{\alpha}})}\delta_E \\ \dot{q} = M_{\dot{\alpha}}\dfrac{Z_u}{(U_1 - Z_{\dot{\alpha}})}u + \dfrac{Z_\alpha}{(U_1 - Z_{\dot{\alpha}})}\alpha - \dfrac{g\sin(\Theta_1)}{(U_1 - Z_{\dot{\alpha}})}\theta + \\ \quad + \dfrac{(Z_q + U_1)}{(U_1 - Z_{\dot{\alpha}})}\dot{\theta} + \dfrac{Z_{\delta_E}}{(U_1 - Z_{\dot{\alpha}})}\delta + \quad + (M_u + M_{Tu})u + \\ \quad + (M_\alpha + M_{T\alpha})\alpha + M_q q + M_{\delta_E}\delta_E \\ \dot{\theta} = q \end{cases} \quad (15)$$

This system of equations can take on the form

$$\begin{cases} \dot{x}_{long} = A_{long} \cdot x_{long} + B_{long} \cdot u_{long} \\ y_{long} = C_{long} \cdot x_{long} + D_{long} \cdot u_{long} \end{cases} \quad (16)$$

where x_{long} represents the set of longitudinal state variables and is given by

$$x_{long} = \begin{bmatrix} u \\ \alpha \\ q \\ \theta \end{bmatrix} \quad (17)$$

and u_{long} is known as the longitudinal input column. If only δ_E is considered as a longitudinal control surface, u_{long} reduces itself to a scalar , $u_{long} = [\delta_E]$.

Therefore, the longitudinal state equations take on the form

$$\begin{bmatrix} \dot{u} \\ \dot{\alpha} \\ \dot{q} \\ \dot{\theta} \end{bmatrix} = A_{long} \cdot \begin{bmatrix} u \\ \alpha \\ q \\ \theta \end{bmatrix} + B_{long} \cdot [\delta_E] \quad (18)$$

The longitudinal state matrix, A_{long} is given by

$$A_{long} = \begin{bmatrix} X'_u & X'_\alpha & X'_q & X'_\theta \\ Z'_u & Z'_\alpha & Z'_q & Z'_\theta \\ M'_u & M'_\varepsilon & M'_q & M'_\theta \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (19)$$

where

$X'_u = (X_u + X_{Tu})$; $X'_\alpha = X_\alpha$, $X'_\theta = -g\cos(\Theta_1)$; $X'_q = 0$;

$X'_{\delta_E} = X_{\delta_E}$; $X'_u = (X_u + X_{Tu})$; $X'_\alpha = X_\alpha$;

$X'_\theta = -g\cos(\Theta_1)$; $X'_q = 0$; $X'_{\delta_E} = X_{\delta_E}$;

$$Z'_u = \frac{Z_u}{(U_1 - Z_{\dot{\alpha}})};\ Z'_\alpha = \frac{Z_\alpha}{(U_1 - Z_{\dot{\alpha}})};\ Z'_u = \frac{Z_u}{(U_1 - Z_{\dot{\alpha}})};$$

$$Z'_\theta = -\frac{g\sin(\Theta_1)}{(U_1 - Z_{\dot{\alpha}})};\ Z'_{\delta_E} = \frac{Z_{\delta_E}}{(U_1 - Z_{\dot{\alpha}})};$$

$M'_u = M_{\dot{\alpha}} Z'_u + M_u + M_{Tu}$; $M'_\alpha = M_{\dot{\alpha}} Z'_\alpha + M_\varepsilon + M_{T\alpha}$;

$M'_\theta = M_{\dot{\alpha}} Z'_\theta$; $M'_q = M_{\dot{\alpha}} Z'_q + M_q$; $M'_{_E} = M_{\dot{\alpha}} Z'_{\delta_E} + M_{\delta_E}$.

The eigenvalues of A_{long} are coincident with the roots of the longitudinal characteristics equation, associated with the short period and the phugoid modes. The elements of the longitudinal output vector y_{long} could be the entire x_{long} along with a number of additional parameters, for example climb angle, altitude and others [2].

III. DYNAMIC ENGINE MODEL

Mechanical systems dynamics due to the rotating inertias constitute the most important contribution to the engine transient behavior. In fact, shaft speeds are directly linked with mass flow through engine and thrust, which is the main output to be manipulated by the propulsion control system. Additional dynamics due to the gas mass storage and heat transfer between gas and metal are present, but their use is reserved to high-fidelity, detailed models. Control design fundamentally deals with the dynamics of a system. Because of the complexity of the gas turbine, the analysis of engine dynamics is multidisciplinary, that is, it includes physical phenomena in mechanical, fluid mechanic, and thermodynamic fields. Shaft dynamics represent the simplest form and the most important dynamic behavior of a gas turbine engine. The acceleration of the rotor (consisting of the compressor, turbine and the shaft) based on the principle of Newtonian mechanics is

$$\dot{\omega} = \frac{\Delta Q}{I} \quad (20)$$

where $\dot{\omega}$ is the angular acceleration of the rotor, $\Delta Q = Q_T - Q_C$ represents the difference between the torque produced by the turbine, Q_T, and the torque required by the compressor, Q_C, and I is the mass moment of inertia of the compressor-shaft-turbine body [3].

The angular velocity ω is substituted by the shaft rotational speed N and the differential torque ΔQ is represented by a function of shaft speed and fuel flow rate W_f. Substituting these in the torque function, the equation for the shaft rotational speed is expressed as

$$\dot{N} = \frac{f(N, W_f)}{I} \quad (21)$$

Engine dynamics arise from complex, interacting phenomena: gas-flow behavior in the compressor and turbine, shaft inertias and losses, fuel flow transport delay, combustion and the thermal behavior of the engine and its surroundings. Due to the intricate geometry of the engine components and the complexity of gas flow, algebraic expression for f is unavailable. This function is highly dependent on external variables such as aircraft speed and atmospheric conditions, which act as its parameters.

The linear model of shaft dynamics for one-spool engine, based on the Taylor's series expansion of the function f at a (steady-state) nominal operating point is

$$\dot{N} = \frac{1}{I}\frac{\partial f}{\partial N}\cdot \Delta N + \frac{1}{I}\frac{\partial f}{\partial W_f}\cdot \Delta W_f \tag{22}$$

The partial derivatives of *f* are obtained from a perturbation method based on experimental, collected from a host of sensors, from which relevant quantities such as adiabatic efficiencies and stall margins are computed. To obtain an approximate value of a partial derivative, the dependent variable is slightly perturbed from its steady value, and the corresponding increment in net torque recorded. While a controls-oriented model can be as simple as a single linear model corresponding to a point in the grid, a high-fidelity simulation model uses partial derivative information obtained from a fine grid covering the entire flight envelope [4].

The output equation for any engine variable *y* can be expressed as a function of speed and fuel flow as well, so that a small variation of the output variable from its nominal value is expressed as

$$y = \frac{\partial y}{\partial N}\cdot \Delta N + \frac{\partial y}{\partial W_f}\cdot \Delta W_f \tag{23}$$

The first order partial derivatives, similar to stability derivatives in airplane dynamics, are evaluated at the specific nominal condition, about which the linear model is expected to apply.

The transfer function from the input variable fuel flow to the output variable *y* is expressed as

$$\frac{Y(s)}{W_f(s)} = \frac{cb}{s-a} + d \tag{24}$$

where

$$a = \frac{1}{I}\frac{\partial Q}{\partial N};\ b = \frac{1}{I}\frac{\partial Q}{\partial W_f};\ c = \frac{\partial y}{\partial N};\ d = \frac{\partial y}{\partial W_f}$$

For a gas turbine engine, the coefficient *a* is always less than zero in the control envelope. Equation (24) represents a first-order lag, which means that the speed response behaves like a lag function after the fuel flow is changed.

The linear model of shaft dynamics for a two-spool jet engine can be derived by extending the one-spool model of equations (21) and (22) with the dynamics of the second shaft. For a two-spool engine we have

$$\begin{cases} \dot{N}_1 = \dfrac{1}{I}\left(\dfrac{\partial Q_1}{\partial N_1}\cdot \Delta N_1 + \dfrac{\partial Q_1}{\partial N_2}\cdot \Delta N_2 + \dfrac{\partial Q_1}{\partial W_f}\cdot \Delta W_f\right) \\ \dot{N}_2 = \dfrac{1}{I}\left(\dfrac{\partial Q_2}{\partial N_1}\cdot \Delta N_1 + \dfrac{\partial Q_2}{\partial N_2}\cdot \Delta N_2 + \dfrac{\partial Q_2}{\partial W_f}\cdot \Delta W_f\right) \end{cases} \tag{25}$$

Similarly, the output equation is given by

$$y = \frac{\partial y}{\partial N_1}\cdot \Delta N_1 + \frac{\partial y}{\partial N_2}\cdot \Delta N_2 + \frac{\partial y}{\partial W_f}\cdot \Delta W_f \tag{26}$$

In the matrix notation, the shaft dynamics for a tow-spool engine are expressed as

$$\begin{bmatrix} \dot{N}_1 \\ \dot{N}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} N_1 \\ N_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} W_f \tag{27}$$

$$y = \begin{bmatrix} c_1 & c_2 \end{bmatrix}\begin{bmatrix} N_1 \\ N_2 \end{bmatrix} + [d]W_f \tag{28}$$

The frequency-domain representation of two-spool engine dynamics expressed in transfer function form for the output variable *y* is

$$\frac{Y(s)}{W_f(s)} = C(sI - A)^{-1}B + D = \frac{k(s+z_1)}{(s+r_1)(s+r_2)} \tag{29}$$

where *I* is the identify matrix. This transfer function represents a second-order dynamic system [5].

The general nonlinear form of the state and the output equations of an engine can be expressed by the following equations

$$\begin{cases} \dot{x}(t) = f[x(t), u(t), t] \\ y(t) = g[x(t), u(t), t] \end{cases} \tag{30}$$

where *f* and *g* are nonlinear functions of the state variable, the input variable and time. For gas turbine engine, *f* and *g* are smooth enough, within the engine's operating envelope, to have a Taylor-series approximation around the nominal operating condition x_0 and u_0, so that

$$\begin{cases} f[x(t), u(t), t] = f[x_0(t), u_0(t), t] + \\ \qquad + \dfrac{\partial f}{\partial x}(t)\Big|_{\substack{x_0 \\ u_0}} \delta[x(t)] + \\ \qquad + \dfrac{\partial f}{\partial u}(t)\Big|_{\substack{x_0 \\ u_0}} \delta[u(t)] + 0(\delta_f^2) \\ g[x(t), u(t), t] = g[x_0(t), u_0(t), t] + \\ \qquad + \dfrac{\partial g}{\partial x}(t)\Big|_{\substack{x_0 \\ u_0}} \delta[x(t)] + \\ \qquad + \dfrac{\partial g}{\partial u}(t)\Big|_{\substack{x_0 \\ u_0}} \delta[u(t)] + 0(\delta_g^2) \end{cases} \tag{31}$$

where $0(\delta^2)$ represents a small-order term and is the collection of the remaining terms in the Taylor-series

expansion. Choosing small enough deltas in state variables, input variables and in time, the $0(\delta^2)$ is negligible because

$$\begin{cases} x(t) = x_0(t) + \delta[x(t)] \\ u(t) = u_0(t) + \delta[u(t)] \end{cases} \tag{32}$$

Equations (30) and (31) can be written as

$$\begin{cases} \dot{x}(t) = \dfrac{\partial f}{\partial x}(t)\Big|_{\substack{x_0 \\ u_0}} \delta[x(t)] + \dfrac{\partial f}{\partial u}(t)\Big|_{\substack{x_0 \\ u_0}} \delta[u(t)] \\ y(t) - y_0(t) = \dfrac{\partial g}{\partial x}(t)\Big|_{\substack{x_0 \\ u_0}} \delta[x(t)] + \dfrac{\partial g}{\partial u}(t)\Big|_{\substack{x_0 \\ u_0}} \delta[u(t)] \end{cases} \tag{33}$$

For a multi-input and multi-output system, as in the case of the gas turbine engine, the linearized equations (33) can be expressed in matrix form as

$$\begin{aligned} \dot{x}(t) = & \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \cdot & \cdot & \cdots & \cdot \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}_{\substack{x_0 \\ u_0}} \cdot \delta[x(t)] + \\ & + \begin{bmatrix} \dfrac{\partial f_1}{\partial u_1} & \dfrac{\partial f_1}{\partial u_2} & \cdots & \dfrac{\partial f_1}{\partial u_m} \\ \dfrac{\partial f_2}{\partial u_1} & \dfrac{\partial f_2}{\partial u_2} & \cdots & \dfrac{\partial f_2}{\partial u_m} \\ \cdot & \cdot & \cdots & \cdot \\ \dfrac{\partial f_n}{\partial u_1} & \dfrac{\partial f_n}{\partial u_2} & \cdots & \dfrac{\partial f_n}{\partial u_m} \end{bmatrix}_{\substack{x_0 \\ u_0}} \cdot \delta[u(t)] \end{aligned} \tag{34}$$

$$\begin{aligned} y(t) = & \begin{bmatrix} \dfrac{\partial g_1}{\partial x_1} & \dfrac{\partial g_1}{\partial x_2} & \cdots & \dfrac{\partial g_1}{\partial x_n} \\ \dfrac{\partial g_2}{\partial x_1} & \dfrac{\partial g_2}{\partial x_2} & \cdots & \dfrac{\partial g_2}{\partial x_n} \\ \cdot & \cdot & \cdots & \cdot \\ \dfrac{\partial g_n}{\partial x_1} & \dfrac{\partial g_n}{\partial x_2} & \cdots & \dfrac{\partial g_l}{\partial x_n} \end{bmatrix}_{\substack{x_0 \\ u_0}} \delta[x(t)] + \\ & + \begin{bmatrix} \dfrac{\partial g_1}{\partial u_1} & \dfrac{\partial g_1}{\partial u_2} & \cdots & \dfrac{\partial g_1}{\partial u_m} \\ \dfrac{\partial g_2}{\partial u_1} & \dfrac{\partial g_2}{\partial u_2} & \cdots & \dfrac{\partial g_2}{\partial u_m} \\ \cdot & \cdot & \cdots & \cdot \\ \dfrac{\partial g_n}{\partial u_1} & \dfrac{\partial g_n}{\partial u_2} & \cdots & \dfrac{\partial g_l}{\partial u_m} \end{bmatrix}_{\substack{x_0 \\ u_0}} \delta[u(t)] \end{aligned} \tag{35}$$

These equations represent the extracted linear model of the engine from a nonlinear simulation.

IV. REDUCED ORDER MODEL

The dynamics of an engine can be represented by a linear, time variant model near a nominal operating point within a finite time interval.

For a detailed modeling of jet-powered aircraft, it might be needed to include the engine dynamics in terms of variation of thrust following a deflection of the throttle. For most cases, the engine response can be modeled using the following transfer function

$$\frac{T(s)}{\delta_T(s)} = \frac{a}{1 + s\,b} \tag{36}$$

where T is the thrust, δ_T is the throttle deflection and (a, b) are essentially constants whose values depend on the characteristics of the propulsion system [6].

Rearranging the above relationship we have

$$T(s)(1 + s\,b) = a\,\delta_T(s) \tag{37}$$

or

$$s\,T(s) = \frac{1}{b}\left(a\,\delta_T(s) - T(s)\right) \tag{38}$$

Taking the inverse Laplace transform we get

$$\dot{T}(t) = \frac{1}{b}\left(a\,\delta_T(t) - T(t)\right) \tag{39}$$

Thus, the following augmented longitudinal state variable model can be introduced featuring T as an additional state variable and δ_T as an additional input

$$\begin{bmatrix} \dot{u} \\ \dot{\alpha} \\ \dot{q} \\ \dot{\theta} \\ \dot{T} \end{bmatrix} = \begin{bmatrix} X'_u & X'_\alpha & X'_q & X'_\theta & 0 \\ Z'_u & Z'_\alpha & Z'_q & Z'_\theta & 0 \\ M'_u & M'_\alpha & M'_q & M'_\theta & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\dfrac{1}{b} \end{bmatrix} \cdot \begin{bmatrix} u \\ \alpha \\ q \\ \theta \\ T \end{bmatrix} + \begin{bmatrix} X'_{\delta_E} & 0 \\ Z'_{\delta_E} & 0 \\ M'_{\delta_E} & 0 \\ 0 & 0 \\ 0 & \dfrac{a}{b} \end{bmatrix} \cdot \begin{bmatrix} \delta_E \\ \delta_T \end{bmatrix} \tag{40}$$

V. NUMERICAL RESULTS

Starting from hypothesis that compressor air flow rate, G_a, is equal to the turbine gas flow rate, G_{gT}, applying the energy and mass conservation theorems, one can get

$$\begin{cases} \dfrac{\pi J}{30}\dfrac{dn}{dt} = M_T - M_C \\ \dfrac{T_2^*}{T_1^*} = 1 + \left(\pi_C^{*\frac{k-1}{k}} - 1 \right) \dfrac{1}{\eta_C^*} \\ G_a = G_{gT} \\ \dfrac{T_4^*}{T_3^*} = 1 - \left(1 - \dfrac{1}{\pi_T^{*\frac{k'-1}{k'}}} \right) \eta_T^* \\ W_f H_u \eta_{ca} = c_p G_a \left(T_3^* - T_2^* \right) \end{cases} \quad (41)$$

where n is the rotational speed, J- inertia momentum, G-flow rate for air and gases (subscripts a and g), H_u - low heating value of fuel, W_f - fuel flow rate, π_C and π_T - pressure ratio for compressor and turbine, T-temperature in the engine sections, c_p - specific heat at constant pressure, k and k' – ratio of specific heats for air and gases, η - efficiency [7].

The set of equations for the two-spool engine has the following form [8]:

$$\begin{cases} (T_1 \cdot s + \rho_1)\overline{n}_1 - k_{T_{41}^*}^{(1)} \overline{T}_{41}^* + k_{p_4^*}^{(1)} \overline{p}_4^* - k_{p_{41}^*}^{(1)} \overline{p}_{41}^* + k_{p_{21}^*}^{(1)} \overline{p}_{21}^* = 0 \\ (T_2 \cdot s + \rho_2)\overline{n}_2 - k_{T_3^*}^{(2)} \overline{T}_3^* - k_{p_2^*}^{(2)} \overline{p}_2^* + k_{p_{21}^*}^{(2)} \overline{p}_{21}^* + k_{p_{41}^*}^{(2)} \overline{p}_{41}^* = 0 \\ \overline{T}_2^* - k_{p_2^*}^{(3)} \overline{p}_2^* + k_{p_{21}^*}^{(3)} \overline{p}_{21}^* = 0 \\ k_{n_1^*}^{(4)} \overline{n}_1 - k_{n2}^{(4)} \overline{n}_2 + k_{p_{21}^*}^{(4)} \overline{p}_{21}^* - k_{p_2^*}^{(4)} \overline{p}_2^* = 0 \\ k_{p_{21}^*}^{(5)} \overline{p}_{21}^* + k_{n2}^{(5)} \overline{n}_2 - k_{p_2^*}^{(5)} \overline{p}_2^* - k_{T_3^*}^{(5)} \overline{T}_3^* = 0 \\ \overline{T}_{41}^* - \overline{T}_3^* - k_{p_2^*}^{(6)} \overline{p}_2^* - k_{p_{41}^*}^{(6)} \overline{p}_{41}^* = 0 \\ \overline{T}_4^* - \overline{T}_{41}^* - k_{p_{41}^*}^{(7)} \overline{p}_{41}^* - k_{p_4^*}^{(7)} \overline{p}_4^* = 0 \\ k_{p_2^*}^{8} \overline{p}_2^* + k_{T_3^*}^{(8)} \overline{T}_3^* - k_{p_{41}^*}^{(8)} \overline{p}_{41}^* - k_{T_{41}^*}^{(8)} \overline{T}_{41}^* = 0 \\ k_{p_{41}^*}^{(9)} \overline{p}_{41}^* + k_{T_{41}^*}^{(9)} \overline{T}_{41}^* - k_{p_4^*}^{(9)} \overline{p}_4^* - k_{T_4^*}^{(9)} \overline{T}_4^* = k_{S_a}^{(9)} \overline{S}_a \\ k_{p_2^*}^{(10)} \overline{p}_2^* + k_{n2}^{(10)} \overline{n}_2 + k_{p_{21}^*}^{(10)} \overline{p}_{21}^* + k_{T_3^*}^{(10)} \overline{T}_3^* - k_{T_2^*}^{(10)} \overline{T}_2^* = k_{W_f}^{(10)} \overline{W}_f \end{cases} \quad (42)$$

The transfer function from the input variables (fuel flow, $\overline{G}_c(s)$, and the exit nozzle area $\overline{S}_a$), to the output variable, LPC rotational speed $\overline{N}_1(s)$ and HPC rotational speed $\overline{N}_2(s)$, for a two-spool engine with

$$G_a = 90\,kg/s;\ T_3^* = 1500\,K;\ \pi_{C1} = 4;\ \pi_{C2} = 7;$$
$$I_1 = 9.9\ kg \cdot m^2;\ I_2 = 11.43\ kg \cdot m^2;$$
$$n_1 = 9000\ rot/min;\ n_2 = 11000\ rot/min;$$

are

$$\overline{N}_1(s) = \frac{(1.3s + 2.95)\overline{G}_c + (3.65s + 6.27)\overline{S}_a}{1.10s^2 + 4.99s + 5.60}$$

$$\overline{N}_2(s) = \frac{(2.51s + 5.87)\overline{G}_c - 3.32\overline{S}_a}{1.10s^2 + 4.99s + 5.60}$$

The responses of the rotational speeds $\overline{N}_1(t) = \Delta N_1 / N_1$ and $\overline{N}_2(t) = \Delta N_2 / N_2$ to an impulse input (Dirac function), to a unit step input (Heaviside function) for $\overline{G}_c$ and $\overline{S}_a$ are presented in Figures 1-7.

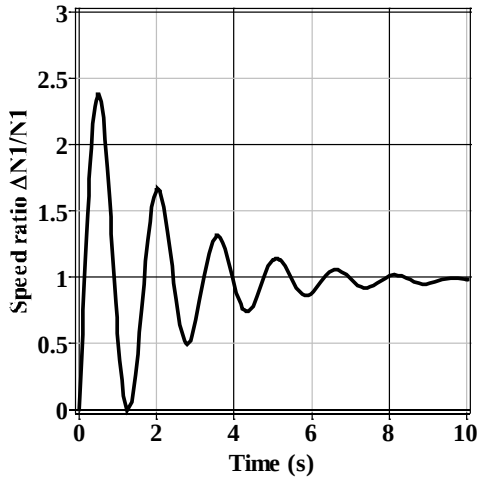


Fig. 1 step response of the LPC rotational speed to the fuel flow rate input

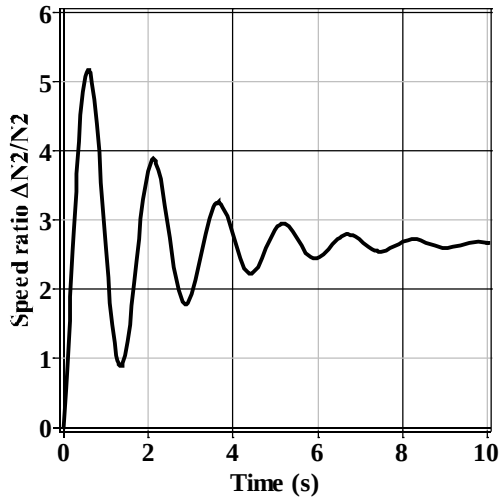


Fig. 2 step response of the HPC rotational speed to the fuel flow rate input

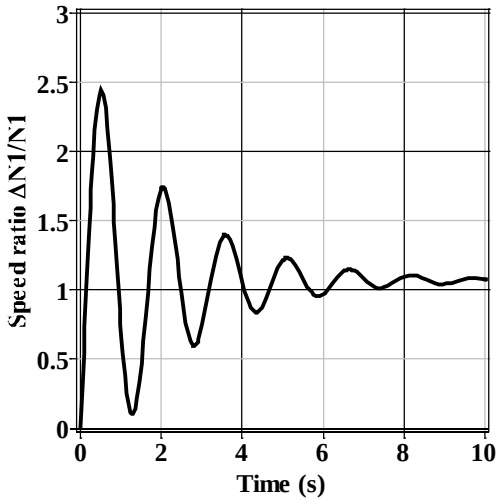


Fig. 3 step response of the LPC rotational speed to the nozzle area input

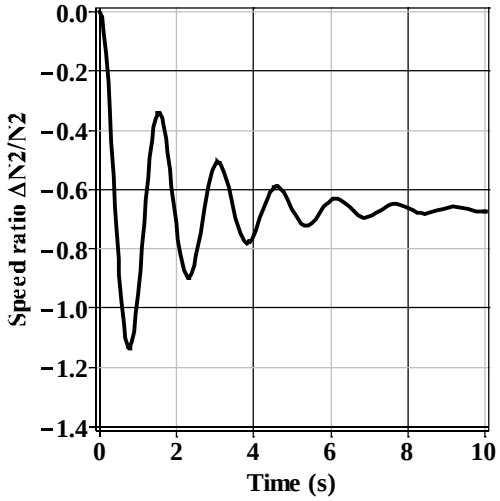


Fig. 4 step response of the HPC rotational speed to the nozzle area input

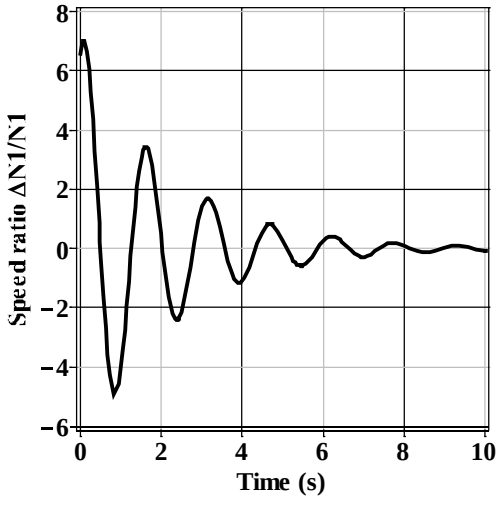


Fig. 5 impulse response of the LPC rotational speed to the fuel flow rate input

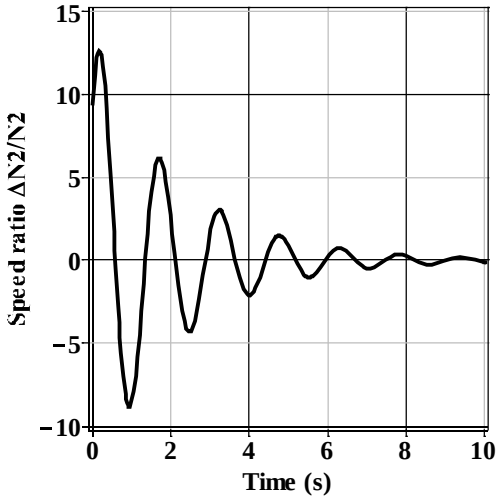


Fig. 6 impulse response of the HPC rotational speed to the fuel flow rate input

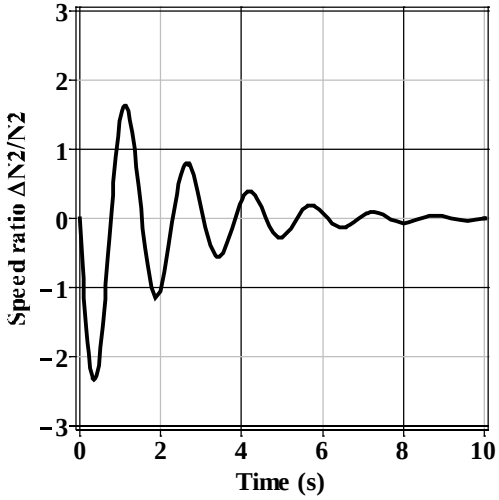


Fig. 7 impulse response of the HPC rotational speed to the nozzle area input

The responses of the aircraft to the throttle input are presented in Figures 8-11.

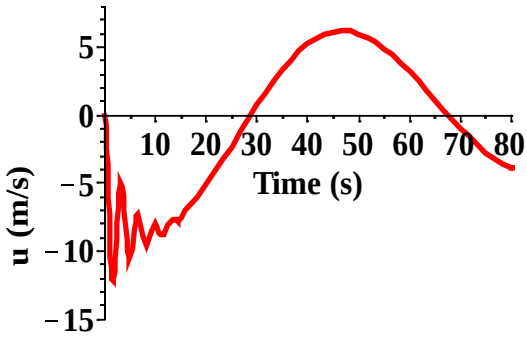


Fig. 8 aircraft velocity impulse response

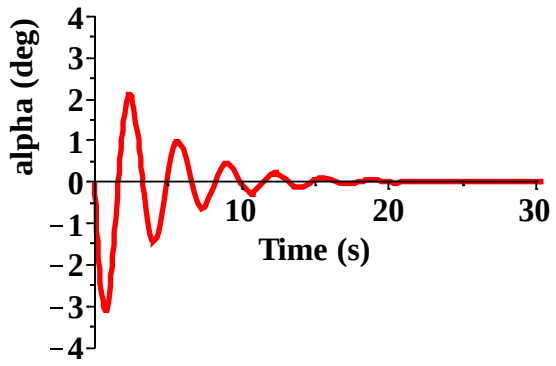


Fig. 9 angle of attack impulse response

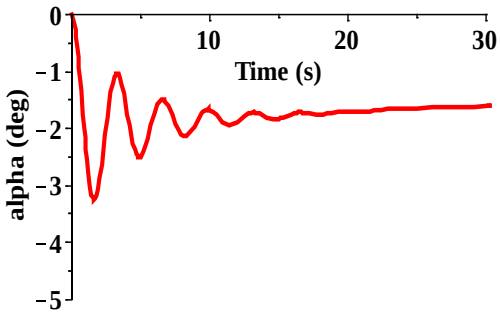


Fig. 10 angle of attack step response

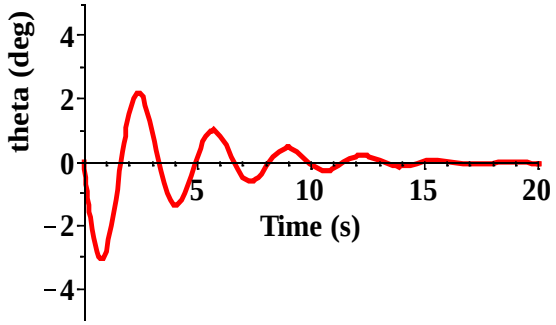


Fig. 11 pitch angle impulse response.

VI. CONCLUSIONS

There are two ways of obtaining faster thrust response: redesigning the regulators for larger closed-loop bandwidths and relaxing the protective limits on variables which tend to peak as thrust response is made faster. Among the variables displaying such peaking are turbine outlet temperature, which peaks during acceleration, stall margin, which tends to undershoot during acceleration and combustor pressure, which tends to undershoot during deceleration.

In order to approach new constructive solutions for complex systems like the aircraft engines, it is necessary to use software codes to study the behavior or performances of these systems. For an aircraft engine the experimental validation is expensive, so the first step in this study is the simulation of the virtual model of constructive elements, like combustion chamber.

The classical linear compensation is adequate only to govern the engine close to a fixed operating point, as defined by the current inlet conditions and desired thrust set point. Aside from nonlinearity and parametric changes in the engine, critical variables must be maintained within safety ranges. Linear compensation is the basic building block of standard aircraft engine control system. Parametric changes and nonlinearity are addressed with gain-scheduled linear compensators while limit protection logic schemes are used to override the active linear regulator when a critical variable approaches its safety limit. Even with constant control gains, limit relaxation is reflected in faster responses, and conversely, the main output response will become slower if limits are made more restrictive.

REFERENCES

[1] M. R. Napolitano, *Aircraft Dynamics. From Modeling to Simulation*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2012.

[2] C. Rotaru, "Nonlinear Characteristics of Helicopter Rotor Blade Airfoils: An Analytical Evaluation," *Mathematical Problems in Engineering*, vol. 2013, Article ID 503858, 9 pages, 2013.

[3] J. D. Mattingly, "*Elements of Propulsion. Gas Turbines and Rockets*", AIAA, Reston Virginia, USA, 2006.

[4] C. Rotaru, P. G. Matei, M. Mihăilă, R. I. Edu, "Dynamic of a turbojet engine considered as aquasi-static system", *International Journal of Mechanics*, ISSN 1998-4448, Volume 8, 2014, pp. 158-166.

[5] C. Rotaru, M. Mihailă, P. G. Matei, R. I. Edu, "Thermodynamic performances of the turbojet combustion chambers – numerical evaluation", *Proceedings of the 2014 International Conference on Mechanics, Fluid Mechanics, Heat and Mass Transfer*, ISBN 978-1-61804-220-0, pp. 86-91.

[6] T. V. Chelaru and M. Cernat, Mathematical model and flight simulation for guided self-supporting gyroplane, *Proceedings of the 9th WSEAS International Conference on Simulation, Modelling and Optimization*, Book Series Mathematics and Computers in Science and Engineering, Pages 401-406, Published 2009.

[7] M. Stoia-Djeska and F. Mingireanu "Mathematical and Computational Model for the Analysis of Micro Hybrid Rocket Motor", *9th International Conference Mathematical Problems in Engineering, Aerospace and Sciences (ICNPAA 2012)*, Book Series AIP Conference Proceedings, Volume 1493, Pages 983-987, Published 2012.

[8] C. Rotaru, A. Arghiropol and C. Barbu, "Some aspects regarding possible improvements in the performances of the aircraft engines", *Proceedings of the 6th IASME/WSEAS International Conference on Fluid Dynamics and Aerodynamics*, Greece, 2008.

Measurement of Boundary Position in Liquid Medium

Yumi Takizawa and Atsushi Fukasawa

***Abstract*—**This paper first gives novel measurement method of liquid surface using both components of transmission and reflection signals simultaneously. This paper then presents a novel measurement scheme composed of the proposed method and conventional chirp method with cascaded connection for fine resolution of positioning against long transmission line. The validation of the proposed method is proceeding using a practical system for transportation of liquid materials.

***Keywords*—**Measurement of liquid surface, transmission and reflection components of signals, the Chirp method, long transmission line of liquid material, cascaded connection.

I. Introduction

MEASURMENT methods of position of liquid surface is required widely in industrial domain for storage, transportation, and production and control of domain of liquid materials.

The chirp method using microwave reflection at liquid surface has been utilized commonly for materials with relatively high reflection at the surface. This method is not applied for materials with relatively low reflection of microwave at the surface.

Recently novel methods are requested for materials with very low reflection of microwaves at the surface. Furthermore fine resolution is requested for measurement of the boundary position between the air and the liquid.

Wave (signal) reflection method is not applied because of small reflection at the surface. Wave transmission method is not used practically, because so many components of signal wave are required to get the solution of boundary position.

This paper first gives novel measurement method using transmission and reflection components signals simultaneously.

This paper then gives a novel measurement scheme composed of the proposed method and conventional chirp method with cascaded connection for fine resolution of positioning against long transmission line.

This system will be used for system of liquid surface positioning in wide and practical areas.

This work was supported in part by the joint research project with Dr. Masaji Abe, COE, Musasino Co., Ltd., and by the trans-disciplinary project by Pro. Hiroe Tsubaki, Vice Director-General, the institute of Statistical Mathematics, Japan.

Yumi Takizawa is with the Institute of Statistical Mathematics, Tachikawa, Tokyo, 190-8562 Japan (phone: 81-50-5533-8539, fax: 81-42-526-4332; e-mail: takizawa@ism.ac.jp).

Atsushi Fukasawa was with Chiba University, Chiba, Japan. He is now with Musasino Co., Ltd., Ota-ku, Tokyo, Japan (e-mail: fukasawafuji@yahoo.co.jp).

II. Estimation of Position of Boundaries

A. Transmission of Electric Signal in Inhomogeneous Medium

In inhomogeneous mediums, relative dielectric constant is assumed as $\varepsilon_r(y)$, which depends on position at neighbor of y. Velocity $c(y)$ at point y is given as follows.

$$c(y) = \frac{c_0}{\sqrt{\varepsilon_r(y)}} \tag{1}$$

where, c_0 is the velocity of light in vacum. Electrical signal trasmission time T is calculated as follows.

$$T(y) = \int_0^y \frac{\eta}{c(\eta)} d\eta \tag{2}$$

B. Signal Transmission of Electrical Signal in Discontinuous Medium

Figure 1 shows configuration of an inhomogenoeous transmission line. A medium boundary (or liquid surface) is included at a certain position of the line.

A measurement system must provide fine resolution for decision of boundary position against long tranmission line length.

A measurement method is given by sectionning of whole span into a boundary section and the other sections without boundary as shown in Fig. 1. The transmission line is devided into N sections. A section k is defined by positions $p_{k-1}$and p_k . A boundary is included at a section n , n is unknown.

Positions $p_{k-1}$and p_k are defined as reflection points. p_0 is transmission point of electrical signal, and p_N is end point of transmission line.

It is considered that some obstercles are inserted at points $p_1 \sim p_{N-1}$ along y axis to yield small reflection of electrical signals.

ISBN: 978-1-61804-253-8

The length of section k is x_k (k = 1, …, n, .. , N). The position of boundary is p_{Bn} in n-th section. δ_n is the distance from p_{n-1} to p_{Bn}.

The passing time t_k in the section k without boundary is given as;

$$t_k = T_k - T_{k-1} = \frac{x_k}{c_k} \quad (3)$$

The velocity c_k in section k is given as;

$$c_k = \frac{x_k}{t_k} \quad (4).$$

where, t_k is passing time in section k.

Now, boundari section is considered.

Passing time t_n of boundary section n is given by velocities c_{n-1} *and* c_{n+1} at the preceeding and the post sections adjacent to the section n.

$$t_n = T_n - T_{n-1} = \frac{\delta_n}{c_{n-1}} + \frac{x_n - \delta_n}{c_{n+1}} \quad (5)$$

$$= \frac{\delta_n\, c_{n+1} + (x_n - \delta_n)\, c_{n-1}}{c_{n-1}\, c_{n+1}} \quad (6)$$

$$= \frac{\delta_n\, (c_{n+1} - c_{n-1}) + x_n\, c_{n-1}}{c_{n-1}\, c_{n+1}} \quad (7)$$

The position of boundary at y_{Bn} is given as follows;

$$y_B = y_{n-1} + \delta_n \quad (8)$$

$$\delta_n = \frac{(T_n - T_{n-1})\, c_{n-1}\, c_{n+1} - x_n\, c_{n-1}}{\{c_{n+1} - c_{n-1}\}} \quad (9)$$

$$= \frac{c_{n-1}\, c_{n+1}}{c_{n+1} - c_{n-1}} \left\{ (T_n - T_{n-1}) - \frac{x_n}{c_{n+1}} \right\} \quad (10).$$

C. Measurement of Boundary Position

(1) Detection of boundary section

Difference of velocity ε_k is considered between adjacent sections.

$$\varepsilon_k = | c_{k-1} - c_k | \quad (11)$$

The value of ε_k is larger than values at the other sections in transmission line.

By iterative calculation of k (1~N), the section n with the maximum value ε_n is given as follows.

$$n = \left\{ k \,\middle|\, \max_k \left(| c_{k-1} - c_k | \right) \right\} \quad (12).$$

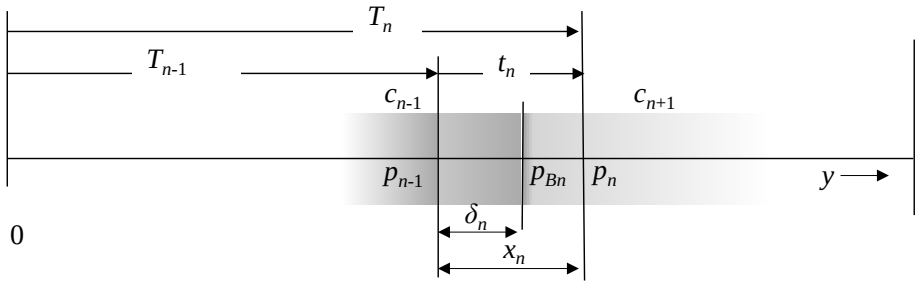


Fig.1 Structure of a section n with a boundary at point p_{Bn}.

.

(2) Measurement of velocities c_{n-1} and c_{n+1}

The adjacent velocities c_{n-1}, c_{n+1} ($n = 1 \sim N$) are considered. Except boundary section n, the velocities of the other sections are equal to c_1 and c_N respectively

$$c_1 = \frac{x_1}{T_1 - T_0} \quad (13)$$

$$c_N = \frac{x_N}{T_N - T_{N-1}} \quad (14)$$

where, T_0 is 0, and c_0 is the velocity of signal in the air.

The time of $T_1 \sim T_N$ are given in experiments.

III. Measurement System for

A. The Chirp method

The transmission times T_k are calculated in chapter II.

Here, these times are obtained by the chirp method.

Chirp signal is made of single carrier modulated by triangle wave. The chirp signal is transmitted and reaches to reflection points p_1, .., pn, .., pN and liquid boundaries in a neuron. The reflection signals from these points and boundaries are received at p_0.

Then the transmission signal s_t and receiving signal s_r are mixed with different frequency. The beat frequency Δf becomes as follows;

$$\Delta f = f_t - f_r = 2\alpha T_C \quad (15)$$

$$\alpha = \{ f_t(\max) - f_t(\min) \} / T_0 \quad (16)$$

where, f_t and f_r are frequencies of transmission and receiving when the signals are mexed. T_C is transmission time from p_0 to reflecting point. α is chirp modulation parameter difined by maximum and minimum frequencies of f_t (max), f_t (min), and T_0 is the time length of triangle wave for chirp modulation.

Then the transmission time T_C is obtained by the following.

$$T_c = \Delta f / 2\alpha \quad (17)$$

Transmission time between reflection points and each boundary is calculated by frequency deviation between transmission and receving signals.

B. System Configuration

A system configuration is shown in Fig. 2. The left part is the chirp method, and the right part is the proposed method, which is connected by cascading to the chirp method.

The part of the chirp method outputs times of reflection components $T_1 \sim T_N$, and T_{Bn}. These times are transmission time from point $p0$ to each reflection point.

The part of the proposed method outputs position of boundary p_{Bn} with fine resolution for input times $T_1 \sim T_N$, and T_{Bn} .

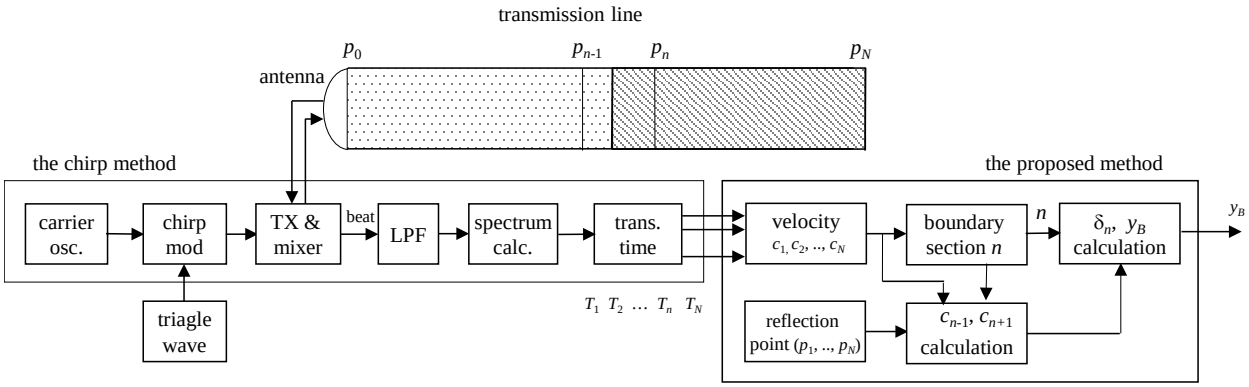


Fig. 2 System configuration of electrical measurement scheme.

IV. Conclusion

This paper presented novel measurement method utilizing transmission and reflection components of signals simultaneously.

This paper then presented a novel measurement scheme composed of the proposed method and conventional chirp method with cascaded connection.

The proposed method is applied to practical liquid transmission equipment under construction.

Acknowledgment

The authors express their sincere gratitude for prof. Toshiharu Horie, Chiba University, associate prof., Kazuhiko Natori, central director of chemotherapy center, Toho University, prof. Alicia Gonzalo-Ruiz, Institute of Neuroscience of Castilla Leon, Spain, and Prof. Hiro-aki Takeuchi, Department of Biology, Shizuoka University, for their helpful advices.

References

[1] Takizawa Y., Fukasawa A., "Electrical Measurement Scheme of Liquid Boundaries in Active Neuron", to be published on *Proc. of Int. Conf. on Health Science and Biomedical Systems (HSBS'14)*, Nov. 22, 2014.

[2] Takizawa Y., Fukasawa A., "Signal Processing by a Neural System and its application to location of multiple events," International Journal of Applied Mathematics and Informatics, Issue 3, vol. 6, pp. 126-133, 2012.

[3] Takizawa Y., Fukasawa A, "Formulation of a Neural System and Modeling of Topographical Mapping in Brain," Proc. of International conference on Circuit, Systems, Control, Signals (CSCS'12) , pp.59 - 64, Barcelona, Spain, Oct. 17, 2012.

[4] Takizawa Y., Fukasawa A., "Formulation of a Neural System and Application to Topographical Mapping," Proceedings of the 3rd International Conference on Neurology (NEUROLOGY '12), pp. 248-253, Kos Island, Greece, July 14-17, 2012.

[5] Takizawa Y., Fukasawa A., "Organization of a Neural System and its Operation for Sensing of Multiple Events in 3D space," Proc. of International Conference on Biomedicine and Health engineering (BIHE'14), pp. 112-118, Tenerife, Spain, Jan. 10, 2014. Tenerife, Spain, Jan. 2013.

[6] Castsigeras E. "Self-synchronization of networks with a strong kernel of integrate and fire excitatory neurons," *WSEAS Transactions on Mathematics*, Issue 7, Vol. 12, pp. 786 – 797, July 2013.

[7] Takizawa Y., Fukasawa A., Formulation of Topographical Mapping in Brain with a Synchronous Neural System, *Proceedings of the 15th International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems (MAMECTIS'13)*, pp. 60–65, Lemesos, Cyprus, Mar. 21-23, 2013.

[8] Fukasawa A. Takizawa Y., Activity of a Neuron and Formulation of a Neural Group for Synchronized Systems, *International Journal of Biology and Biomedical Engineering*, Issue 2, vol. 6, pp. 149-156, 2012.

[9] Takizawa Y., Fukasawa A., Formulation of a Neural System and Analysis of Topographical Mapping in Brain, *International Journal of Biology and Biomedical Engineering*, Issue 2, vol. 6, pp. 157-164, 2012.

[10] Fukasawa A., Takizawa Y., Activity of a Neuron brought by Electro-Physical Dynamics, *International Journal of Mathematical Models and Methods in Applied Sciences*, Issue 8, Volume 7, pp. 737-744, 2013.

[11] Takizawa Y., Fukasawa A., Electrical Measurement Scheme of Liquid Boundaries in Active Neuron, to be published in *Proc. of International Conference on Health Science and Biomedical Systems*, Nov. 22, 2014.

[12] Fukasawa A., Takizawa Y., Aactivities of Neuron and Unicellular Organism for Positive Pulse generation, to be published on *Proc. of Int. Conf. on MMCTSE'14*, Nov. 28, 2014.

Yumi Takizawa. Yumi Takizawa received the B.S. degree in Physics from Shinshu University in 1984, and the Ph.D. degree from the University of Tokyo in 1994. She joined the Institute of Statistical Mathematics (ISM) as an associate professor in 1995. She received the Prize on Telecommunication System Technology from the Foundation of Telecom Association, Japan in 2004. She has been engaged in neural systems in brain based on electro-physical and biological studies at the University of Virginia, USA and the ISM, Japan. She has been awarded for the Best Paper on NEUROLOGY'12.

Atsushi Fukasawa received the Master of Arts degree and the Ph.D. degree from Waseda University in 1967 and 1983. He joined Graduate School of Natural Science, Chiba University as a professor in 1997. He received the Award of the Agency of Science and Technology, Japan in 1982, and Ohm (publisher) Prize in 1994. He received Telecommunication System Technology Prize from the Foundation of Telecommunication Association, Japan in 2004. He is a senior member of the IEEE. He has been awarded for the Best Paper on NEUROLOGY'12.

The Gravitational Constant G from the standpoint of Quantum Vacuum Dynamics and Polarizable – Vacuum Approach to General Relativity

Luigi Maxmilian Caligiuri

***Abstract*—** The ultimate physical understanding of the force of gravity and its unification with the other three fundamental forces are still missing. To this aim important insights can stem from the study of the meaning of the gravitational constant G. In this paper, starting from ZPF inertia hypothesis and the Polarizable-Vacuum approach to General Relativity, a novel model of G as a function of quantum vacuum energy density is proposed. In particular, on this basis it has been argued that G could actually be a function of the distance from the mass generating the gravitational field whose analytical expression, under the hypothesis of spherically symmetric body and slowly varying gravitational potential, is derived. Finally, an eventual experimental test of the model, making use of precise measurements of G performed on a satellite, orbiting around the Earth, has been outlined. The proposed model apart from giving new insights into a deeper understanding of gravitation could set the basis for surprising applications related, for example, to the field of gravity control and space propulsion.

***Keywords*—** Gravitation; General Relativity; Quantum Vacuum; Velocity of Light; Vacuum Polarization; Standing Waves; Planck Scale; ZPF Inertia Hypothesis.

I. Introduction

GRAVITY is the most mysterious and still incompletely understood among the fundamental forces of Nature. A reason for this could probably derives from the description that General Theory of Relativity (GTR) gives of it in terms of "spacetime metric" which may hide some fundamental underlying physical details.

To this aim the study of the physical origin of the gravitational constant G ruling the strength of gravitation, through the well know Newton's law of universal gravitation

$$\vec{F}_g = G\left(m_1 \cdot m_2 \big/ r^2\right)\vec{u}_{\vec{r}} \tag{1}$$

where m_1 and m_2 are the interacting masses and $\vec{r}$ is their relative distance vector, could give very important insights. In the commonly accepted theoretical framework G is generally assumed to be a universal constant whose value is assumed to be [1]

$$G = \left(6.67384 \pm 0.00080\right)\cdot 10^{-11}\, m^3 \cdot kg^{-1} \cdot s^{-1} \tag{2}$$

regardless of the magnitude of the mass generating gravitational field and of the distance from it. Nevertheless, since from the beginning of the past century, some interesting models, involving a variable gravitational constant G has been proposed by authoritative scientists [2,3,4].

A very interesting and intriguing suggestion, in part also coming from these studies, is that G could be somehow related to the QED Zero – Point – Field (ZPF) or the so-called Quantum Vacuum (QV).

Indeed ZPF is the only "medium" between gravitational matter and then the relation between (QV) and gravitation is a task of primary importance in order to establish a quantum theory of gravity. Several theoretical and experimental results have shown QV can be influenced by electromagnetic fields [5,6,7]. In addition to electromagnetic field, also the presence of matter is though able to modify the structure of QV. In 1967, Sakharov [8] suggested gravity could be the effect of a change in the quantum-fluctuation energy of ZPF quantum vacuum induced by the presence of matter as experimentally demonstrated by the Casimir effect [9,10,11].

Later, starting from Sakharov's results, Puthoff [12] proposed the hypothesis that ordinary matter could be ultimately composed of sub-elementary constitutive charged entities he called "partons", able to dynamically interact with the fluctuating QED quantum vacuum according to a sort of resonance mechanism. According to Puthoff's model, the inertia of a body would be the result of the interaction between partons and ZPF quantum fluctuations whose effect would result in the modification of the electromagnetic modes of ZPF at the interface between a body and its surrounding space determining the so – called Zero-Point-Field Lorentz force [13]. In this way both the inertial and gravitational masse of a body could be substantially composed of confined e. m. modes of ZPF whose presence modifies the previous state of QED QV.

On the other hand, it is a known fact, theoretically explained within the GTR and supported by strong experimental evidences, that the gravitational potential generated by mass, depending on the radial distance from it, affects the running rate of clocks, the measure of distances as well as the velocity of light.

All the above results and many other available in the literature [14,15,16] strongly suggest QV can be actually considered as a sort of "optical" medium equipped with a own inner structure.

Luigi Maxmilian Caligiuri is with Foundation of Physics Research Center (FoPRC), Celico, CS 87053 ITALY and University of Calabria, Arcavacata di Rende, CS 87036 ITALY (phone: +39094431875; fax: +390984431875; e-mails: caligiuri@ foprc.org; max.caligiuri@gmail.com).

On this basis, Puthoff [17] showed that, under "standard" (weak-field) astrophysical conditions, the basic principles of GTR can be coherently reformulated in terms of the changes in the permittivity ε_0 and permeability μ_0 constants of a polarizable vacuum (PV) of an optical electromagnetic medium whose properties are able to reproduce all the features of GTR under the above conditions.

Moreover, in some recent works [18-21] the author proposed a model of (QV), characterized by a Planckian metric, described in terms of the dynamics of energy density in which inertial and gravitational mass are interpreted as local stable variations of QV energy density with respect its "unperturbed" value. Within this model, gravity is interpreted as originated by the local gradients of QV energy density $\Delta\rho(\vec{r},t)$, due to presence of mass, giving an unbalanced ZPF pressure that manifest itself as gravitational force.

More recent researches [8,22,23] also suggested the possibility that G could be truly expressed as a function of more fundamental physical quantities, i.e. the so – called "Quantum Vacuum Zero Point Field Mass – Density Equivalent" ρ_{QV} giving a measure of the energy density of the QED QV and the Planck time t_P.

In this paper, starting from the interpretation of inertial and gravitational mass as the seat of standing waves of ZPF [13,23] and from the picture of QV as a special optical medium characterized by a refraction index [17,24], a novel model describing the gravitational constant G as a function of QV energy density, has been proposed.

these result to be compatible with those already obtained by the author in a previous work [25], starting by different hypotheses.

In particular, the model results suggest the gravitational "constant" G could be not truly unchanging but actually varies as a function of the radial distance from the mass originating gravitational potential itself and whose approximate analytic expression, for a spherically symmetric mass distribution and a weak and slowly varying (with distance from mass) gravitational potential, has been discussed.

Finally, an eventual experimental test of the model, making use of precise measurements of G on a satellite orbiting around the Earth (or a generic celestial body of known mass), has been synthetically sketched. The present results are compatible with those already obtained by the author in a previous work [25], starting from different hypotheses.

II. Inertial and Gravitational Mass as the Result of Stading Waves of Quantum Vacuum

In the model of Hairsh, Rueda and Puthoff (HRP) [12,13] a material body is represented, with respect to the electromagnetic interaction, as a resonant cavity in which a suitable set of oscillating modes of QV. According to this hypothesis (also known as ZPF Inertia Hypothesis), the inertial and gravitational masses m_i and m_g associated to a given material body are given by

$$m_i = m_g = \left(V_0/c^2\right)\int_0^\infty \eta(\omega)\rho(\omega)\,d\omega \tag{3}$$

in which ω is the angular frequency of ZPF mode, $\rho(\omega)$ is the spectral energy density of quantum vacuum ZPF fluctuations and $\eta(\omega)$ is a function that would quantify the fraction of ZPF energy density that electromagnetically interacts with the particles contained in the "useful volume" V_0 or, in other words the "efficiency" of interaction [13]. In this way the apparent inertial mass of a given object would originate by the interaction, during the accelerated motion of the body, between the ZPF energy density fraction enclosed in the object (given by $\eta(\omega)$) and the partons contained in the volume V_0.

The equivalence between the apparent inertial mass m_i and passive gravitational mass m_g (namely the "weak" equivalence principle) then "naturally" arises, within the HRP model, from the consideration that a body accelerating through ZPF is identical to a body that remains fixed in a gravitational field and having the QED QV fall past on curved geodesic.

It is now important to stress the physical meaning of such volume, that must intended as an electromagnetic resonant cavity with conducting wall, being the volume V_0 the space enclosed within the cavity walls. In this way the electromagnetic modes of ZPF are trapped inside the cavity and the resulting energy is accumulated in it.

This energy cannot be however accumulated without limit, since the possible electromagnetic modes inside a resonant cavity are upper bounded by a limiting frequency ω_{up} whose value is substantially determined by the plasma frequency ω_{Pl} of the electrons in the cavity walls.

The connection between the modes inside cavity with those outside it is allowed by the conductive structure of cavity walls.

Now if we consider an ideal resonant cavity (i.e. neglecting energy dissipation of modes) at the absolute temperature $T = 0$, outside the cavity there are only the ZPF quantum fluctuations while inside it there is a discrete number of possible modes possible oscillating at their exact characteristic frequencies ranging from 0 up to ω_{Pl}.

So, said N the maximum number of this modes, we have

$$E_{tot} = \sum_{k=1}^{N} \hbar\omega_k/2 \tag{4}$$

where $\omega_1 \le \omega_2 \le \ldots \le \omega_N \le \omega_{Pl}$. Now, under the above assumptions, the energy given by (4) must be equal to the quantity given by (3) multiplied by c^2, namely

$$E_{tot} = mc^2 = V_0\int_0^\infty \eta(\omega)\rho(\omega)\,d\omega = \sum_{k=1}^{N} \hbar\omega_k/2 \tag{5}$$

On the other hand we know that the density of ZPF electromagnetic oscillation modes in the frequency interval between ω and $\omega + d\omega$ is given by

$$N(\omega)d\omega = \left(\omega^2 d\omega / \pi^2 c^3\right) d\omega \tag{6}$$

and, assuming an average energy per mode equal to $\hbar\omega/2$, we obtain the spectral energy density of ZPF fluctuation as

$$\rho(\omega)d\omega = \left(\hbar\omega^3 / 2\pi^2 c^3\right) d\omega \tag{7}$$

that substituted into (5) gives

$$\eta(\omega) = \sum_{k=1}^{N} \left(\pi^2 c^3 / V_0\right) \left[\delta(\omega - \omega_k) / \omega_k\right] \tag{8}$$

Equation (8) states that the spectrum of electromagnetic field inside the cavity is composed by a sum of N lines placed at $\omega = \omega_k$ whose amplitude diminishes with the increase of frequency.

Equation (8) holds under the simplification that no dissipation occurs. Nevertheless it can be shown [13] that, if the dissipation is small, a more accurate expression for the line-shaped functions $\delta(\omega - \omega_k)/\omega_k$ is given by the so-called Lorentzian –lineshape function, whose expression is

$$l(\omega) = \left(\Delta\omega / 2\pi\right) \left[\left(\omega - \omega_0\right)^2 + \left(\Delta\omega / 2\right)\right]^{-1} \tag{9}$$

where the quantity $\Delta\omega$ is the lineshape broadening parameter and describes the various types of dissipation and broadening effects.

By discretizing (9) and using it in (8), we have

$$\eta(\omega) = \left(\pi^2 c^3 / 2\pi\omega V_0\right) \sum_{k=1}^{N} \Delta\omega_k \left[\left(\omega - \omega_k\right)^2 + \left(\Delta\omega_k / 2\right)\right]^{-1} \tag{10}$$

where, as above, ω_k is the proper frequency of the k-th mode and $\Delta\omega_k > 0$ its frequency broadening.

Finally, the mass associated to a resonant cavity (not including the overall mass of the walls) is given by (3) using the result of (10), namely

$$\begin{aligned} m &= \left(V_0 / c^2\right) \int_0^\infty \eta(\omega)\rho(\omega)d\omega = \\ &= \int_0^\infty \left(\pi^2 c^5 / 2\pi\omega\right) \rho(\omega) \times \\ &\times \sum_{k=1}^{N} \Delta\omega_k \left[\left(\omega - \omega_k\right)^2 + \left(\Delta\omega_k / 2\right)\right]^{-1} d\omega \end{aligned} \tag{11}$$

now using (7) in (11) and recalling the definition of energy given by (5), we can write

$$\begin{aligned} m &= \int_0^\infty \left(\pi^2 c^5 / 2\pi\omega_0\right) \rho(\omega) \times \\ &\times \sum_{k=1}^{N} \Delta\omega_k \left[\left(\omega - \omega_k\right)^2 + \left(\Delta\omega_k / 2\right)\right]^{-1} d\omega = \\ &= \int_0^\infty \left(\pi^2 c^5 / 2\pi\omega_0\right) \left(\hbar\omega^3 / 2\pi^2 c^3\right) \times \\ &\times \sum_{k=1}^{N} \Delta\omega_k \left[\left(\omega - \omega_k\right)^2 + \left(\Delta\omega_k / 2\right)\right]^{-1} d\omega = \\ &= \left(1/c^2\right) \sum_{k=1}^{N} \int_0^\infty \left(\Delta\omega_k / 2\pi\right) \left[\left(\omega - \omega_k\right)^2 + \left(\Delta\omega_k / 2\right)\right]^{-1} \times \\ &\times \left(\hbar\omega / 2\right) d\omega \end{aligned} \tag{12}$$

or, in a more compact form,

$$m = \sum_{k=1}^{N} \int_0^\infty A_k(\omega) \left(\hbar\omega / 2\right) d\omega \tag{13}$$

where we have posed

$$A_k(\omega) = c^{-2} \int_0^\infty \left(\Delta\omega_k / 2\pi\right) \left[\left(\omega - \omega_k\right)^2 + \left(\Delta\omega_k / 2\right)\right]^{-1} \tag{14}$$

Equation (13) is very meaningful since it shows the total mass inside the resonant cavity associated to a body can be expressed, even in the presence of dissipation, as the overlapping of the zero point energies of all the electromagnetic modes of QV, each of them broadened by a suitable factor given by (14). Furthermore, it is expected that the most part of modes are not overlapping as long as the cavity size remains small, since their frequency separation will become comparable with the broadening $\Delta\omega$ only at the highest frequencies [13].

III. The Effect of Gravitational Field in the Polarizable-Vacuum Approach to General Relativity

In the Polarizable – Vacuum (PV) formulation of GTR [17], the metric changes in terms of variations of permittivity and permeability of a polarizable QV and the Maxwell's equations in curved space are treated in the isomorphism of a polarizable medium characterized by a variable refractive index in flat space. In this way, as already suggested by Eddington [26] almost a century ago, the bending of a light ray around a massive object could be considered as a refraction effect of the space (actually the vacuum) in flat space. In this model the reduction of light velocity (as well as all the other effects on time and length intervals) are interpreted as the effect of an effective increase of the refractive index of QV.

The basic assumption of PV approach is to consider that the presence of a mass induces vacuum polarization effects so that the polarizability of vacuum around a mass differs from its asymptotic value (in the far-field condition). Formally it is done by assuming, for the vacuum, the electric flux density $\vec{D}$ is given by the following expression

$$\vec{D} = \varepsilon\vec{E} = K\varepsilon_0\vec{E} \tag{15}$$

where $\vec{E}$ is the electric field, ε_0 is the permittivity of vacuum interpreted as the polarizability of the vacuum per unit of volume, and K is the modified (by the presence of mass) dielectric constant of the vacuum (considered as a general function of position) due to the vacuum polarizability changes. The quantity K then represents, within PV model, the fundamental variable since it rules the variations of all fundamental physical quantities due to the altered properties of medium (QV) in the presence of the mass. Some theoretical cosmological considerations about fine structure constant [15] require

$$\varepsilon\left(K\right) = K\varepsilon_0, \quad \mu\left(K\right) = K\mu_0 \tag{16}$$

then, accordingly, the light velocity will be a function of K

$$c\left(K\right) = c/K \tag{17}$$

where c is the asymptotic light velocity in flat space ($K = 1$). Equation (17) has a very important meaning since it shows the dielectric constant of vacuum is a sort of refractive index of the PV in which vacuum polarizability changes in response to GTR-induced effects. Equation (17) implies a "rescaling" of the other fundamental physical quantities as energy, mass, time and length intervals. In particular, as observed by Dicke [15] by using a limited principle of equivalence, if E_0 is the energy of a system in a flat space ($K = 1$), we have, in general

$$E = E_0/\sqrt{K} \tag{18}$$

in a region of space in which $K \neq 1$. The combined use of (17) and (18) gives

$$m = K^{\frac{3}{2}}m_0 \tag{19}$$

for the rest mass m_0 of a particle. As a consequence of a change in energy due to the variation of QV polarizability we have, starting from $E = \hbar\omega$,

$$\omega = \omega_0/\sqrt{K} \tag{20}$$

and the correspondent equations for time interval

$$\Delta t = \Delta t_0\sqrt{K} \tag{21}$$

and for length interval

$$\Delta l = \Delta l_0/\sqrt{K} \tag{22}$$

According to (22), the dimension of a material object varies with the local changes in QV polarizability so reproducing, from a different standpoint, the variable metric of GTR. Furthermore, according to (21) and (22), the "natural" measurement of light velocity by rods and clocks returns the unperturbed value c so maintaining the invariance of the locally measured velocity of light assumed by Einstein's Theory of Relativity.

The key point, for our following treatment, is the explicit expression of K as a function of the mass and the distance from it. It can be obtained by following a lagrangian approach [27]. The starting point is the lagrangian of a free particle

$$L = -mc^2\sqrt{1 - \left(v/c\right)^2} \tag{23}$$

that, using (17), becomes

$$L = -\left(m_0c^2/\sqrt{K}\right)\sqrt{1 - \left(Kv/c\right)^2} \tag{24}$$

whose density $\overline{L}$ is

$$\overline{L} = -\left(m_0c^2/\sqrt{K}\right)\sqrt{1 - \left(Kv/c\right)^2}\,\delta^3\left(\vec{r} - \vec{r}_0\right) \tag{25}$$

being $\vec{r}$ the generic position and $\vec{r}_0$ is the position of the particle with respect a given frame or reference. The interaction between a particle of charge q and an electromagnetic field given by the four-potential $\left(\phi, \vec{A}\right)$ is described by the lagrangian density $\overline{L}_p$

$$\overline{L}_p = -\left[\left(m_0c^2/\sqrt{K}\right)\sqrt{1 - \left(Kv/c\right)^2} + q\phi - q\vec{v}\cdot\vec{A}\right]\times \\ \times\delta^3\left(\vec{r} - \vec{r}_0\right) \tag{26}$$

On the other hand the lagrangian density of the electromagnetic field itself is, in PV formulation

$$\overline{L}_{em} = -\left(1/2\right)\left(B^2/K\mu_0 - K\varepsilon_0E^2\right) \tag{27}$$

and we must also specify a lagrangian density for the quantity K, being considered as a scalar variable, that can be obtained, by imposing the standard Lorentz-invariant form for the propagational disturbance of a scalar and following [15], in the form

$$\vec{L}_K = -\left(c^4/32\pi GK^2\right)\times \\ \times\left[\left(\nabla K\right)^2 - \left(K/c\right)^2\left(\partial K/\partial t\right)^2\right] \tag{28}$$

so that the total lagrangian density for a matter-field interaction in a PV with a variable dielectric constant, is

$$\overline{L} = \overline{L}_p + \overline{L}_{em} + \overline{L}_K \tag{29}$$

The equation of motion in a dielectric vacuum is then obtained, using the principle of least action, by the variation of the lagrangian density $\delta\int\overline{L}\,dVdt$ with respect the particle variables $\left(x, y, z, t\right)$ while the equation, in K, describing the effect of vacuum polarization due to the presence of matter and field, is get by the variation of $\overline{L}$ with regard to the K variable

$$\nabla^2\sqrt{K} - \left(K/c\right)^2\left(\partial^2\sqrt{K}/\partial t^2\right) = \\ = -\left(8\pi G\sqrt{K}/c^4\right)\times \\ \times\left\{\left(m_0c^2/2\sqrt{K}\right)\left[\left(1 + \left(Kv/c\right)^2\right)/2\sqrt{1 - \left(Kv/c\right)^2}\right]\delta^3\left(\vec{r} - \vec{r}_0\right)\right\} + \\ -\left(4\pi G\sqrt{K}/c^4\right)\left(B^2/K\mu_0 + K\varepsilon_0E^2\right) + \\ +\left(\sqrt{K}/4K^2\right)\left[\left(\nabla K\right)^2 + \left(K/c\right)^2\left(\partial K/\partial t\right)^2\right] \tag{30}$$

We are now interested in the simplified case of a static field, for which $\partial K/\partial t = 0$, with spherical symmetry. In this case we have, from (30), after some simple manipulations

$$d^2\sqrt{K}/dr^2 + \left(2/r\right)\left(d\sqrt{K}/dr\right) = \left(1/\sqrt{K}\right)\left(d\sqrt{K}/dr\right)^2 \tag{31}$$

whose solution, satisfying the Newtonian limit, is [17]

$$K = \left(\sqrt{K}\right)^2 = e^{2GM/rc^2} \tag{32}$$

As shown in [17] this solution correctly reproduces the usual GTR Schwarzschild metric in weak-field conditions as those occurring in the Solar System and is in agreement with the scaling factor used in the previous analysis proposed by this author [25].

IV. The constant G as a function of quantum vacuum energy density in the Polarizable-Vacuum model

A. The relation between gravitational constant G and QV energy density

As well-known, the physical vacuum cannot be considered, due to Heisenberg uncertainty principle, as a void deprived by any physical dynamics but as physical entity manifesting a complex and fundamental background activity in which, even in absence of matter, processes like virtual particle pair creation – annihilation and electromagnetic fields fluctuations, known as zero point fluctuations (ZPF) continuously occur.

The maximum amount of "virtual" energy density $\rho_{QV,\max}$ stored in the "unperturbed" ZPF fluctuations of QV can be estimated by considering the Planck's constants. Planck showed, basing on dimensional arguments, that the values of gravitational constant G, velocity of light c and Planck's constant $\hbar$, it was possible to derive some natural units for length, time and mass, i.e. respectively the co-called Planck's length (l_P), time (t_P) and mass (m_P). Then he (and we with him) reversed the point of view by considering these quantities as the most elementary ones, from which the "fundamental" constants (as G, c and so on) can be derived.

In order to assume GTR to remain valid up the Planck scale, we must have

$$\rho_{QV,\max} = m_P c^2 \Big/ l_P^3 \tag{33}$$

where $l_P = 1.616 \times 10^{-35}\, m$, $m_P = 2.177 \times 10^{-8}\, kg$ when G has the currently accepted constant value $6.67384 \times 10^{-11}\, m^3 kg^{-1} s^{-2}$ and $c = 299792458\, m\, s^{-1}$.

The value of $\rho_{QV} \approx 10^{113}\, Jm^{-3}$ so obtained by (33), can be considered as the maximum possible value $\rho_{QV,\max}$ of QV energy density, since it would represent, within the currently accepted picture, the maximum energy density can exist "without being unstable to collapsing space-time fluctuations" [13] associated to the value given in [1].

As already shown [22,23,25] the relationship between the gravitational "constant" G and QV energy density ρ_{QV} can be expressed in a "natural" way by noting that, dimensionally

$$\left[G\right] = \left[L\right]^3 \left[M\right]^{-1} \left[T\right]^{-2} \tag{34}$$

and

$$\left[\rho_{QV}\right] = \left[M\right]\left[L\right]^{-3} \tag{35}$$

where we'll indicate from now on, for simplicity, with ρ_{QV} the so-called Mass – Density – Equivalent (MDE) of QV energy density (equal to ρ_{QV}/c^2 where ρ_{QV} is the originally defined QV energy density function) referring to it simply as QV energy density, so we can write

$$G = 1\Big/\left(\rho_{QV} t_P^2\right) \tag{36}$$

where t_P is the Planck's time whose value is $t_P = 5.391 \times 10^{-44}\, s$ in correspondence to $G = 6.67384 \times 10^{-11}\, m^3 kg^{-1} s^{-2}$.

We can then assume that also G is a function of ZPF energy density defining a fundamental property of space itself, originated from QV and related to the most elementary units for time, length and mass by the equation

$$G = l_P^3 \Big/ \left(m_P c^2 t_P^2\right) \tag{37}$$

Equation (36) can be naturally generalized to the case of a variable QV energy density by formally assuming

$$G\left(\rho_{QV}\right) = 1\Big/\left(\rho_{QV} t_P^2\right) \tag{38}$$

Equation (38) also means that, far from any mass, where the quantum vacuum energy density reaches its "unperturbed" value given by (33), the gravitational constant G, given by (38), takes the value given by (37) while, in the proximity of a mass its value varies according to (38). We'll see in the following that its value, at a given point of space will, depend upon its radial distance from the center of mass of the body (or the system of bodies) generating the gravitational field itself.

B. Quantum Vacuum energy density in a gravitational potential according to the Polarizable-Vacuum approach

The results so far obtained allow us to interpret the mass of a body as the place of occurrence of electromagnetic standing – waves of ZPF that determines a storing of electromagnetic energy density within the body itself.

This dynamics of ZPF together with the consideration of other theoretical elements [13] show that, inside the portion of space associated to "electromagnetically useful" volume V_0, the energy density of ZPF reduces giving rise to a standing wave structure in which this energy is "stored". Outside this structure, on the contrary, the QV energy density is higher and determines the gravitational potential associated to that mass.

This view is also coherent with the model already developed in previous works [23,25,28] in which the inertial mass of a body or particle is interpreted as the result of the reduction of the local QV energy density determining, in its neighborhoods (where the QV energy density is higher), an energy density gradient $\Delta\rho\left(\vec{r},t\right)$ which originates the gravitational potential.

The decrease of energy density inside the massive body can be mathematically proved within the model of PV by considering the correspondence between the parameter K and a refraction index n of QV and calculating the expression of vacuum refraction index inside and outside the massive body for a not too strong gravitational field.

In the model so far discussed the energy spectrum related to ZPF modes of standing waves inside the resonant cavity originating the inertia of a body is substantially discrete and includes a finite number of modes whose frequencies are in the interval $\omega_1 \le \omega \le \omega_N$, with $\omega_N \le \omega_{Pl}$. Nevertheless, when the size of cavity increases so do the number of modes and, due to broadening of frequencies, we obtain a continuous-like frequency spectrum. Physically this must be the case since, when the maximum size of the cavity $L_{\max} \to \infty$, all the modes are possible and we obtain the limit of continuum.

We can then interpret the standing waves inside the resonant cavity associated to a massive body like those generated within a continuum elastic fluid medium [13,23,25,28] whose properties characterize the QV behavior.

Actually, the observation that light propagation in vacuum can be modified by the interaction with an electromagnetic field strongly indicates vacuum itself is a special kind of optical medium. Furthermore, the results of the PV approach to GR, showing a deep analogy QV and a dielectric medium, further indicate this vacuum must have an inner structure that can change by the interaction with matter and electromagnetic fields.

This general view is also consolidated within the QFT by the consideration that all the elementary particles are actually QV excitations and by some recent studies, based on this assumption, picturing space-time as arising from a sort of large-scale condensate of more fundamental objects, in which matter is a collective excitations of these constituents, describable by hydrodynamics techniques [18,29].

For such a medium the relation between the (longitudinal) wave propagation velocity v and the medium density ρ can be written as

$$v = \Gamma \Big/ \sqrt{\rho} \tag{39}$$

where Γ is a constant related to the specific medium characteristics.

By putting our analogy in (39) we have

$$c = \Gamma_{QV} \Big/ \sqrt{\rho_{QV}} \tag{40}$$

where c is the velocity of light, ρ_{QV} is the ZPF energy density and Γ_{QV} a constant related to QV structure. By inserting (40) into (17) and squaring both the members we obtain

$$\rho_{QV} = K^2 \rho_{QV,0} \tag{41}$$

where ρ_{QV} is the QV density in a generic point of the space around the mass, $\rho_{QV,0}$ the asymptotic density in the flat space and K the QV "refraction" index given by (32). It shows that QV energy density in the space surrounding the body is multiplied by a factor K^2 with respect its "unperturbed" asymptotic value.

V. The Constant G as a Function of Distance from the Mass Generating Gravitational Potential

We are now in position to obtain the dependence of gravitational constant G on the distance from the mass originating the gravitational potential.

By using the definition of K, given by (16), in (38) we obtain

$$\rho_{QV}(r) = e^{4GM/rc^2} \rho_{QV,0} \tag{42}$$

where we have explicitly expressed the functional dependence of ρ_{QV} on r and assumed the value of $\rho_{QV,0}$ as constant. Multiplying the (42) side by side by t_P^2, taking the reciprocals and using (38), we obtain our main result

$$G(r) = G_0 e^{-4G(r)M/rc^2} \tag{43}$$

where $G_0 \equiv 1/\rho_{QV,0} t_P^2$. A similar expression for ZPF density can be also obtained by using the (38) in (42), namely

$$\rho_{QV}(r) = e^{4M/r\rho_{QV}(r)c^2 t_P} \rho_{QV,0} \tag{44}$$

Equations (43) and (44) respectively describe the dependence of G and ρ_{QV} on the radial distance r from the mass M generating the gravitational potential.

They are transcendent equations and cannot be solved analytically but their qualitative behavior can be discussed in the case of weak and slowly-varying gravitational fields. In this case we can expand the exponential factor in (43) obtaining, at the first order in G

$$e^{-4GM/rc^2} = 1 - 4GM/rc^2 + \ldots \tag{45}$$

Using this result in (43) we find

$$G(r) = G_0 \left[1 - 4MG(r)/rc^2\right] \tag{46}$$

Equation (46) is a first order approximate equation for $G(r)$ that can be immediately solved to give the solution

$$G(r) = G_0 / \left(1 + 4G_0 M/rc^2\right) \tag{47}$$

The asymptotic behavior of this function appears to be coherent with general physical assumptions since we have

$$\lim_{r \to +\infty} G(r) = G_0 \qquad \lim_{r \to 0} G(r) = 0 \tag{48}$$

the case $r \to +\infty$ corresponding to a point far from gravitational source (in which G assumes its "unperturbed" value G_0), while the case $r \to 0$ to point at the center of spherical symmetrical object in which, as known, gravitational field is zero.

VI. Discussion

A. General considerations

According to the model proposed in this paper, gravity is originated by ZPF energy density gradients due to the presence, at a given point of space, of a massive body. Inside a body, the ZPF energy density is decreased to give rise to the standing waves structure described by (34), and correspondingly increased outside the resonant cavity represented by the body itself, thus generating the gravitational potential associated to that specific body. The increment of QV energy density around a mass decreases with the distance from its center as shown by (44).

When two massive bodies are close each other, the ZPF energy density increase between them is smaller so originating gravitational attraction.

An important remark concerns the physical meaning of the quantity G_0: it represents the value of G at a point "infinitely" far from mass M in which the ZPF is unperturbed. Its value should be determined by experimental measurements (far from any mass) or extrapolated by means of the know value of gravitational field at a given distance from a mass M.

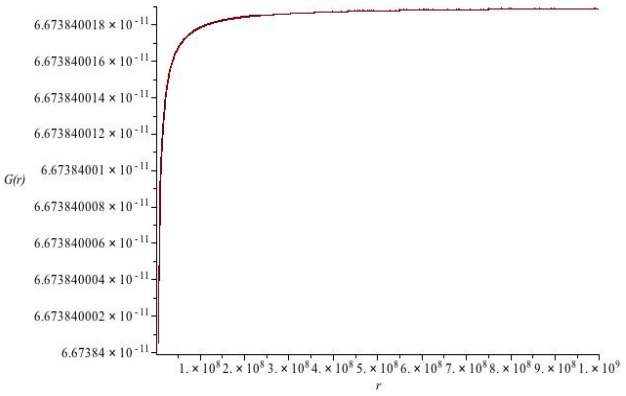


Fig. 1 plot of G vs distance from Earth center for $r \geq R_{Earth}$. The values of G and r are expressed in standard units.

Contrary to what one could think, the value of G_0, within the proposed model, is not equal, in principle, to the quantity $l_P^3 \big/ m_P c^2 t_P^2$ with the Planck units given by the commonly accepted values because the latter in turn are calculated by assuming the value of G given in [1] (namely measured at Earth's surface or deduced by astronomical observations [30] in the presence of massive bodies). Furthermore, we should also consider the contributions to ZPF, and then eventually to G, coming from strong and weak interactions, at this stage not still included in our model of the function describing $G(r)$ (and not considered in the HPR model of inertia as well [12,13]).

It is remarkable to note that (47), like the more general (43), doesn't contain any Planck's unit, allowing, in principle, the calculation of G using only the value of G_0, experimentally determined.

A possible estimation of G_0 could be obtained by using the know value of G at Earth's surface, as given by (1), in the (43) with $r = R_{Earth}$ and $M = M_{Earth}$. Following this procedure we obtain

$$\begin{aligned} G_0 &= G\left(R_{Earth}\right) \times \\ &\times \exp\left[4G\left(R_{Earth}\right) M_{Earth} \big/ R_{Earth} c^2\right] = \\ &= 6.673840019 \times 10^{-11}\, m^3 \cdot kg^{-1} \cdot s^{-1} \end{aligned} \tag{49}$$

where we have assumed $c = 299792458\, m \cdot s^{-1}$, $R_{Earth} = 6372.7955 \times 10^3\, m$ and $M_{Earth} = 5.9736 \times 10^{24}\, kg$.

The value of G_0, given by (49) would represent the asymptotic value of gravitational constant calculated by considering the gravitational field generated by the Earth as if it should be far from all the other masses of Universe.

By using this value of G_0 we can plot, by way of qualitative example, the function $G(r)$ given by (47) as a function of the radial distance r (Fig. 1) from the Earth center.

In evaluating this graphic we must remember that (44) just represents an approximation, at first order, of the value of $G(r)$ that is valid when $\Delta G \to 0$, so it doesn't necessarily represent the actual numerical values of the "whole" function $G(r)$ in particular in the slope-region of the curve, since it is just here the contribution of the higher order terms of the series of (45) could be more important; nevertheless it gives correct indications about the asymptotic behavior of the function $G(r)$ at $r \to 0$ and $r \to +\infty$.

We must note the numerical value given by (49) is slightly higher, on the average, of that commonly assumed [1] in agreement with the predictions of our model. This difference is very tiny since it appears at the eight decimal place so making it difficult to be experimentally revealed by a direct measurement.

This difficulty is also in part due to the need to perform such measurement far from any mass able to influence the results or, equivalently, within a region characterized by a distribution of masses able to nullify, with a very high precision, the gravitational field at the measurement point.

Nevertheless, to this regard, we should recall the estimation given by (49) is based on the simplifying assumption (45) and that the consideration of the higher order terms in the expansion (45) could be able to modify the value of (46), so making the difference $\left|G_0 - G\left(R_{Earth}\right)\right|$ larger and then more easy to be measured.

Furthermore, even if this difference will be confirmed to be so small for a giving mass, this shouldn't imply the

resultant physical effects to be negligible in the proximity of a system of massive bodies, because of the summation of contribution of each of them to the overall altered value of G.

A variable gravitational constant G could have very deep consequences on the current framework of theoretical physics with from GTR to Quantum Field Theory (QFT) and cosmology and unthinkable possible applications in intriguing fields as, for example, the gravity modification and space propulsion, as already suggested by many authors [2,3,4,15,22,28].

B. About a possible experimental test

In principle, a possible experimental test of (43) could be performed by measuring (for example on a satellite, orbiting around the Earth), with very high precision, at specified distances $r_{SAT,i}$ from the Earth center, the values of $G\left(r_{SAT,i}\right)$, comparing the latter with that measured at the Earth's surface $G\left(R_{Earth}\right)$. More specifically we can write $\forall i$, using (43)

$$G\left(R_{SAT,i}\right) = G_0 e^{-4G\left(r_{SAT,i}\right)M/r_{SAT,i}c^2} \tag{50}$$

and

$$G\left(R_{Earth}\right) = G_0 e^{-4G\left(R_{Earth}\right)M/R_{Earth}c^2} \tag{51}$$

Dividing side by side (50) by (51) we obtain the equation

$$\begin{aligned} &G\left(r_{SAT,i}\right)/G\left(R_{Earth}\right) = \\ &= Exp\left\{-\left(4M/c^2\right)\left[G\left(r_{SAT,i}\right)/r_{SAT,i} - G\left(R_{Earth}\right)/R_{Earth}\right]\right\} \end{aligned} \tag{52}$$

in which only measured quantities and known constants appear and that con be then experimentally verified.

In any case, in a realistic set-up, very high measurement precisions would be required in order to reveal the very small variations between the values of $G\left(r\right)$ expected by (52) at different distances from Earth.

Furthermore, the influence of other celestial bodies (firstly the Sun and the Moon) on the ZPF energy density at the measurement points (and consequently on the overall gravitational potential at the same points) should be taken into account. This would introduce additional terms into (43) able to modify the form and behavior of solutions and represents an important theoretical question to be addressed in the future developments of the model, already in progress.

Other questions to be addressed regard the extension of the discussed model to generic shape bodies and the calculation of numerical solutions of (43) able to quantify the contribution of higher order terms to the overall value of $G\left(r\right)$.

VII. Conclusion

In this paper a novel model of gravity, based on the variability of gravitational constant G, expressed as a function of ZPF energy density, has been proposed. Starting from some previous theoretical results, the inertial and gravitational mass of a body have been interpreted as the result of the seat of standing waves of ZPF analogous to longitudinal waves generated inside a continuum elastic medium.

These waves are able to alter the local QV energy density determining a decrease of ZPF energy density within the massive bodies and, correspondingly, a ZPF energy density increment in their surrounding space so generating QV energy density gradients (unbalanced ZPF pressure) giving rise to the gravitational force.

We have also seen that QV can be pictured, within the PV approach, to a dielectric polarizable medium, characterized by a refraction index as function of the gravitational field, in which all the modifications induced by the presence of mass, as described by GTR, can be viewed as due to the altered value of the above refraction index.

Under the above assumptions we have shown that Quantum Vacuum energy density around a spherically symmetric massive body depends on the radial distance from its center and that gravitational constant G also depends on ZPF energy density and on radial distance from the mass generating the gravitational field.

The main result of this work is then a model suggesting that the gravitational "constant" G is actually variable and its dynamics is ruled by the QV energy density, depending on the distance from massive objects.

Under some simplifying assumptions, an approximate analytical expression of the function $G\left(r\right)$, describing the radial dependence of gravitational "constant", has been obtained.

Finally, a possible experimental test of the first model results of the model, involving the accurate measurements of G at different distance from Earth surface, is suggested.

Although the above theoretical model is still in a preliminary phase and involves some simplifying assumptions to be addressed in its future developments, its theoretical, experimental and applicative consequences could be very deep. They will be discussed in details in future and forthcoming publications.

References

[1] CODATA, "*Internationally recommended values of the Fundamental Physical Constants*", 2010.

[2] P. A. M. Dirac, "The cosmological constants", *Nature*, vol. 139, pp. 323-323, February 1937.

[3] R. H. Dicke, "New research on old gravitation: are the observed physical constants independent of the position, epoch, and velocity of the laboratory ?", *Science*, vol. 138, pp. 653-664, March 1959.

[4] J. W. Moffat, S. Rahvar, "The MOG weak field approximation and observational test of galaxy rotation curves", arXiv: 1306.6383.

[5] N. Ahmadi, M. Nouri-Zonoz, "Quantum gravitational optics: effective Raychaudhuri equation", *Phys. Rev. D*, vol. 74, 044034, August 2006.

[6] G. L. J. A Rikken, C. Rizzo, "Magnetoelectric anisotropy of the quantum vacuum", *Phys. Rev. A*, vol. 67, 015801, January 2003.

[7] A. Arnoni, A. Gorsky, M. Shifman, "Spontaneous Z2 symmetry breaking in the orbifold daughter of n=1 super-Yang-Mills theory, fractional domain walls and vacuum structure", *Phys. Rev. D*, vol. 72, 105001, November 2005.

[8] A. D. Sakharov, "Vacuum fluctuation in curved space and the theory of gravitation", *General Relativity and Gravitation*, vol. 32, pp. 365-367, 2000.

[9] H. Gies, K. Klingmuller, "Casimir effect for curved geometries: proximity-force-approximation validity limits", *Phys. Rev. Lett*. 96, 220401, June 2006.

[10] S. K. Lamoreaux, "Demonstration of the Casimir Force in the 0.6 to 6 μm range", *Phys. Rev. Lett*., vol. 78, January 1997.
[11] H. B. Chang et al., "Quantum mechanical actuation of microelectromechanical systems by the Casimir force", *Science*, vol. 291, pp. 1941-1944, March 2001.
[12] H. E. Puthoff, "*Gravity as a zero-point-fluctuation force*", Phys. Rev. A, vol. 39, pp. 2333-2342, 1989.
[13] B. Haish, R. Rueda, H.E. Puthoff, "*Inertia as zero-point-field Lorentz force*", Phys. Rev. A, vol. 49, pp. 678-694, 1994.
[14] H. A. Wilson, "An electromagnetic theory of gravitation", *Phys. Rev*., vol. 17, January 1921.
[15] R. H. Dicke, "Gravitation without a principle of equivalence", *Rev. Mod. Phys*., vol. 29, July 1957.
[16] F. de Felice, "On the gravitational field acting as an optical medium", *Gen. Rel. & Grav*., vol. 2, pp. 331-345, 1971.
[17] H. E. Puthoff, "Polarizable-vacuum (PV) approach to general relativity", *Found. of Phys*., vol. 32, n. 6, pp. 927-943, June 2002.
[18] L. M. Caligiuri, "The emergence of space-time and matter: entropic or geometro-hydrodynamic process ? A comparison and critical review", *Quantum Matter*, vol. 3, n. 3, pp. 249-255, 2014.
[19] L. M. Caligiuri, A. Sorli, "Relativistic energy and mass originate from homogeneity of space and time and from quantum vacuum energy density", *Am. J. of Mod. Phys*., vol. 3, n. 2, pp. 51-59, 2014.
[20] L. M. Caligiuri, A. Sorli, "Gravity originates from variable energy density of quantum vacuum", *Am. J. of Mod. Phys*., vol. 3, n. 3, pp. 118-128, 2014.
[21] L. M. Caligiuri, A. Sorli, "Space and time separation, time travel, superluminal motion and big bang cosmology", *Journal of Cosmology*, vol. 18, pp. 212-222, August 2014.
[22] L. M. Caligiuri, T. Musha, "Quantum vacuum energy, gravity manipulation and the force generated by the interaction between high-potential electric fields and ZPF", *J. of Astrophysics and Space Science*, vol. 2, n. 1, pp. 1-9, 2014.
[23] L. M. Caligiuri, "The not so constant gravitational constant G as a function of quantum vacuum energy density", to appear in Proceedings of the *9th Vigier Conference on Unified Field Mechanics*, Morgan State University, 16-19 November 2014, Baltimore, MD USA.
[24] YE Xing-Hao, L. Qiang, "Inhomogeneous vacuum: an alternative interpretation of curved spacetime", *Chin. Phys. Lett*. Vol. 25, n. 5, 2008.
[25] L. M. Caligiuri, "Gravitational constant G as a function of quantum vacuum energy density and its dependence on the distance from mass", *J. of Astrophysics and Space Science*, vol. 2, n. 1, pp. 10-17, 2014.
[26] A. S. Eddington, "*Space, time and gravitation*", Cambridge: Cambridge University Press, 1920, p. 109.
[27] H. Goldstein, "*Classical Mechanics*", Reading, MA: Addison-Wesley, 1957, pp. 206-207.
[28] R. R. Hatch, "Gravitation: Revising both Einstein and Newton", *Galilean Electrodynamics*, vol. 10, n. 4, pp. 69-75, July/August 1998.
[29] S. Liberati, L. Maccione, "Astrophysical constraint on Planck scale dissipative phenomena", arXiv:1309.7296v1, September 2013.
[30] J. Mould, S. A. Uddin, "Constraining a possible variation of G with Type Ia Supernovae", arXiv:1402.1534v2, February 2014.

Stochastic Response Surface Methodology in Medicine with censored/uncensored data analysis

Teresa Oliveira[1], Conceição Leal[2], Amílcar Oliveira[3]

***Abstract*—** Stochastic Response Surface Methodology (SRSM) is very well known as a tool for modeling uncertainty in simulation models, mainly in the context of Risk Analysis. Developments on SRSM were remarkable over the past few years, mainly due to the technological advances and new powerful tools in what concerns hardware and software. It has also been possible to extend and explore this methodology in many areas and to state its importance in studies connected with Health problems, more specifically in Medicine revealing to be crucial on helping specialists to attain better and accurate Diagnosis and Prognosis. In recent papers, see [4], [5] and [34], [35], the authors paid attention to the application and behavior of the methodology when applied to a sample with censored data. In this work the authors present an extension of previous research involving the analysis of the whole available data contained in Wisconsin Prognostic Breast Cancer database. Uncensored data and right censored together data will be studied in order to compare both censored/uncensored results and to attain more realistic results. The aim is to provide more appropriate models to prediction and to risk analysis, since the models obtained with uncensored data suffer from a right bias. Computational results and graphics were implemented using R software.

***Keywords*—** Breast Cancer Prognosis, Polynomial Chaos, Risk Analysis, Stochastic Response Surface Methodology, Uncertainty.

I. Introduction

In this paper we review Stochastic Response Surface Methodology as a tool for modeling uncertainty in the context of Risk Analysis, specifically in healthcare. An application in the survival analysis in the breast cancer context, with noncensored and censored data, is implemented with R software.

Uncertaity quantification on Risk Analysis has a crucial relevance, particularly in complex dynamical systems. Stochastic process variability and epistemic uncertainty state uncertainty quantification as a prerequisite on probabilistic risk assessment, either in experimental data or in model parameters. Thanks to the recent technological advances, simulation techniques became current to estimate models allowing to predict systems' behaviors, with respect to the probability of occurrence of a specific event and the consequences of this occurrence.

Probabilistic uncertainty quantification demand numerical simulation whose computational costs are often very high, thus the use of metamodels arises as a pressing necessity. The Response Surface Methodology (RSM) is known to be a suitable tool for the quantification of uncertainty. Stochastic Response Surface Methodology (SRSM) was developed as a conceptual extension of the traditional RSM, to approximate model inputs and outputs in terms of random variables, such as standard normal variables, by a polynomial chaos expansion.

The objective of the methodology is to reduce the number of model simulations required for adequate estimation of uncertainty, as compared to conventional Monte Carlo methods. The use of metamodel in conjunction with Monte Carlo method allows a faster estimation of the response system CDF that characterize the uncertainty propagation on the model. The importance of such methodology is very well known, mainly in the survival analysis context, so we will review it in the next section.

II. The Stochastic Response Surface Methodology – Expansion into Polynomial Chaos

Monte Carlo method is the most popular one to probabilistic quantification of covariate uncertainty propagation in the system response and to estimate its probability distribution. However, this may have very high computational costs and it is necessary to rely on methodologies that converge more quickly to the solution.

The SRSM allows the generation of a metamodel, computationally less demanding and statistically equivalent to the complete numerical model, which can be used to implement sensitivity analysis. It is needed a limited number of simulations of the complete model to get the response of the system for the model coefficients estimation. Reference [32] introduce this methodology and present two case studies. The basic idea of the methodology is to represent the response of a model to changes in covariate, using a response surface defined with a polynomial basis that is orthogonal with respect to a probability measure on the space of parameters. The SRSM lies on the assumption that the random variables, whose probability density functions are square integrable, can be approximated by the expansion in stochastic series of random variables or direct transformation of these, see [26] for details.

In the last years SRSM was subject to important developments that release of assumptions that are not always realistic. The arbitrary and data-driven approaches of the methodology are important examples of these developments. In the classic version, a vector of random variables $\xi=(\xi_i)$, $i = 1,..,n$, is selected, under $N(0,1)$ distribution, representing

uncertain variables of a model in such way that $x_i = h(\xi_i)$. After this selection, response variables are represented as a function of the same vector of random variables: $Y = f(c, \xi)$, being c a vector of coefficients to estimate. The response of the system model to the various achievements of ξ, allows the estimation of model coefficients. The coefficients c_i quantify the dependence of the response Y on the input vector ξ, for each realization of x.

Consider Ψ_i polynomials which form a base of orthogonal polynomials to a given probability measure. The form of the function f result of the polynomials chaos expansion is expressed by:

$$Y = f(c,\xi = \\ = c_0\Psi_0 + \Sigma_{i1=1}^{\infty} c_{i1}\Psi_1(\xi_{i1}) + \Sigma_{i1=1}^{\infty}\Sigma_{i2=1}^{i1} c_{i1i2}\Psi_2(\xi_{i1},\xi_{i2}) + \ldots$$

This series is approximate by a truncated polynomial in n-th power. The roots of the polynomial of $n+1$ degree can be used in the simulation model in order to get the data to estimate the model coefficients.

The methodology is implemented sequentially as follows: (i) representation of uncertain input variables; (ii) representation of the response variable; (iii) estimation of the model parameters; (iv) calculation of the response statistical properties; (v) evaluation of the model answers approximation.

The following algorithm illustrates the application of the methodology.

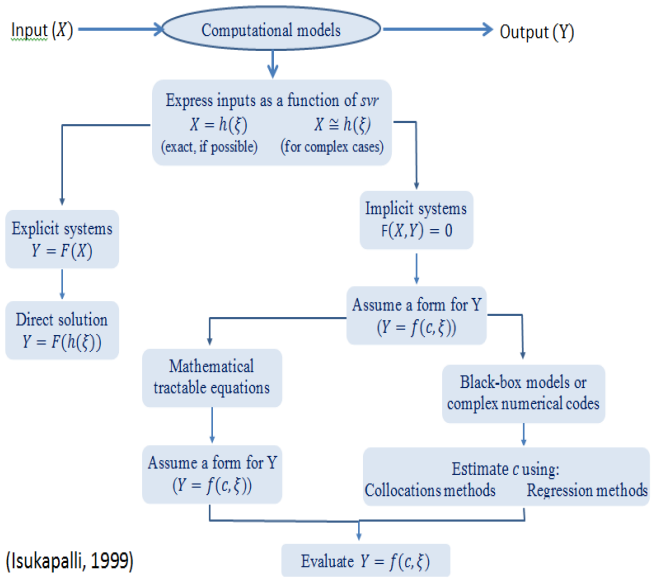


Fig. 1: Schematic depiction of the Stochastic Response Surface Method

In the classical approach the measure is the Gaussian and the polynomials are the Hermite polynomials, which form is as follow (see [19] and [24]).

$$PCE = c_0 + \Sigma_{i=1}^{n} c_i\xi_i + \Sigma_{ii=1}^{n} c_{ii}\left(\xi_i^2 - 1\right) + \Sigma_{i=1}^{n-1}\Sigma_{j>i}^{n} c_{ij}\xi_i\xi_j$$

Reference [10]–[13] showed that it is possible to obtain a better approximation of the response variables using non-Gaussian expansions in polynomials chaos. In this case, the Hermite polynomials are replaced by orthogonal polynomials with respect to the probability measure of input variables, [10]. This approach was designated generalized polynomial chaos expansion. Reference [20] presented conditions on the probability measures involving the mean square convergence of the generalized polynomials chaos expansion.

Reference [30] proposed a new generalization of the methodology, called arbitrary polynomial chaos expansion or data-driven chaos expansion. In this new approach, the probability distributions and the probability measures are arbitrary. Statistical moments are the only source of information that is propagated in the stochastic model. Probability distributions may be discrete, continuous, or continuous discretized, may be specified by analytical way (PDF or through CFD), numerically using a histogram or by using the raw data. In this approach, all distributions are admissible for the input variables of a given model, as long as they have in common a finite number of moments. Thus, in the case of considering truncated polynomial, there is just the need to know a finite number of moments, with no need for complete knowledge of the probability density function or even its existence, which frees the researcher from the need of assumptions that may not always be supported by existing data. According to the literature it is known that this expansion converges exponentially and faster than the classical expansion.

The estimation of model parameters depends on the model complexity, see [33]**.** In case the model is invertible, the parameters can be obtained directly from the input random variables $(\xi_i)_{i=1}^{n}$. If the model equations are mathematically manipulated, in spite of nonlinearities, then the model coefficients can be obtained afterwards, by an appropriate norm minimization of residuals, replacing the input random variables by the respective transformations in terms of Gaussian variables $N(0,1)$, known as Galerkin method, for details see [33]. When the model equations are difficult to manipulate, the coefficients can be estimated by the collocation points method. Each set of points is chosen such that the model estimates are accurate at these points, given by the set of the N resulting linear equations.

Reference [33] present some methods for parameter estimation, all based on the collocation points methods: Probabilistic Collocation Method, Efficient Collocation Method and Regression Based Method and these authors discuss advantages and disadvantages for each method.

The expansion in polynomials chaos is a simple but powerful tool for stochastic modeling. Probability density function, probability distribution functions or other statistics of interest can be estimated and quickly evaluated via Monte Carlo simulation (MCS), once the evaluation of a polynomial function is faster than the original equations model evaluation.

ISBN: 978-1-61804-253-8

In the case of risk analysis, to use arbitrary expansion, one can directly use a set of large-sized data or probability density function of maximum or relative minimum entropy, since, in this case, the relevant moments of the polynomial chaos expansion are compatible with those of the input variables. The bootstrap resampling method may be used to obtain more precise estimates of the moments from a reduced set of data available, providing a more accurate estimation of the risk assessment model. Reference [31] propose such an application on calibration models to history matching for CO_2 storage in underground reservoirs.

III. Literature applications

The RSM in its various approaches has an important role in generating reduced models to replace the simulator complex processes requiring a very high number of simulations. Applications are diverse and many of them relate to the Stochastic Response Surface Methodology for quantification of uncertainty, in stochastic processes.

Reference [32] applied the SRSM to four case studies cover a range of applications, both from the perspective of the model application (biology, air quality and groundwater) and its complexity.

Reference [6] applied the SRSM to study the role of various hydro geologic parameters in the uncertainty assessment of the chemical contaminant concentration in groundwater, resulting from the nuclear industry, so that one can designing the waste disposal facilities and remedial action plans. This study provides a program of environmental monitoring in the nuclear industry.

Reference [8] applied the SRSM to analyze the reliability of stochastic stability of rock slopes involving correlated non-normal variables.

Reference [29] applied the SRSM, based on the expansion of arbitrary polynomial chaos, to a problem of contaminant transport in 3D heterogeneous aquifer and the risk to human health caused by exposure of the population.

References [27]– [31] have applied the SRSM, with different approaches and combined with other methodologies, in various problems relating to the storage of CO_2 in underground geological formations and the associated risks.

Reference [1] combined subset simulation method with the SRSM to perform probabilistic analysis of a shallow strip footing. As well as other aspects, they analyzing uncertainty propagation of the soil shear strength parameters on the ultimate bearing capacity. This problem, that involves the computation of the ultimate bearing capacity of a strip footing, is presented to demonstrate the efficiency of the proposed procedure by comparing the results given by MCS methodology applied on Polynomial Chaos Expansion (PCE) models and on original deterministic model.

Reference [14] applied the SRSM in a model of gas injection into an incompressible porous media, and showed its effectiveness in uncertainty and sensitivity analysis of complex numerical models.

Reference [2] applied the generalized SRSM to assess leakage detectability at geologic carbon storage sites under parameter uncertainty. They demonstrated how it can be used to construct probability maps for assessing the detection of anomalies in the coverage of underground geological storage formations, in space and time.

Reference [7] applied the SRSM to estimate the propagation of uncertainties in the parameters of the function retention of strontium in the human body.

Reference [9] applied the SRSM to analyze the reliability of an underground cavern, associated with deterministic finite element methods. More specifically, the SRSM was used to perform the probabilistic analysis serviceability performance of the cavern.

SRSM was not yet conveniently explored in applications on the field of medicine, despite uncertainty confound the understanding of the essential medical facts and how they are integrated ([25]. The uncertainty parameters, the heterogeneity of the patient and the stochastic uncertainty results are increasingly important concepts in models of medical decision. [25] and [3] show various methods for analyzing uncertainty and the heterogeneity of patients in decision models.

The breast cancer prognosis is markedly heterogeneous, and research has often focused on the effect of prognostic factors related to the disease, such as the expression of estrogen receptors, tumor size and others. Still remains an open question of modeling the data recurrence time, the complexity of the shape of the hazard function over follow-up period, and the identification of factors that may affect it, using a fully parametric approach ([16].

In the next section we present a study in the prognostic breast with uncensored and uncensored data using SRSM.

IV. Application in human health: censored/uncensored data

SRSM was applied to the Wisconsin Breast Cancer Prognostic (WPBC) database which is constituted by data of 253 patients with breast cancer, who have undergone surgical excision of invasive cancer. First, we used only 69 of these cases that correspond to the patients that had remission till the end of the study. Second, we used all cases, including 184 that are right censored because they withdrew from the study or we know only the time of the last examination. The database is very well described by [36], who are responsible for its design and construction.

We estimated, by regression on sample data, a first and a second order polynomial chaos expansion with Hermite

polynomials, in which the response variable is the time to recur (TTR) and the uncertainty parameters are three of the ten available covariates, related with morphological characteristics of the nucleus of malignant cells: AREA, TEXTURE and SIZE. The first two variables result from the average of the measurements in the three nuclei with higher values from all nuclei retained in the image. Since there are more than 30 cases the normality of covariates was assumed, as well as independence once there is a low correlation. Different models were simulated with respect to the degree of the polynomial and to the number and nature of the included covariates. The second order model was used to estimate the distribution of the response variable and the resulting survival and risk functions.

The proposed study was implemented by using the R program language with *survival*, *fitdistrplus*, *Hmisc* and *EQL* packages.

The distribution which best fits to TTR data with remission was gamma distribution (lower values of statistics, AIC and BIC criteria obtained with *gofstat* function in PDF than what we got with *fitdist* function, booth from *fitdistrplus* package).

A data-frame was constructed containing the variables processed in accordance with appropriate transformation to a normal distribution: $x_i = \mu_i + \sigma_i \, \xi_i$.

In the first order model, it was observed that variable SIZE did not give a significant contribution to the model. Adjusting the second order model, it was observed that the intersect and the terms $AREA$, $AREA^2 - 1$ and interaction WAREA× $SIZE$ are significant. A Cox model was fitted and it was observed again that SIZE covariate did not reveal statistical significance. The assumption test to proportional hazards reveals no statistical evidence that these are not proportional to each variable in the overall model.

A similar study was carried out with the censored data.

The distribution that best fits to TTR with censored data was Lognormal distribution (lower values of AIC and BIC criteria obtained with *fitdistcens* function from f*itdistrplus* package).

The Cox model reveal that AREA and SIZE were statistically significant and the second order model shown that the $AREA$, SIZE, $AREA^2 - 1$, $SIZE^2 - 1$ and AREA× $TEXTURE$ interaction are statistically significant.

Using the second order models the response variable PDFs were simulated by Monte Carlo sampling (Fig.2 –data only with remission and censored data) and it can be compared with empirical PDF (Fig.3).

```
> Nsim<-10^4
> t<-0
> X1<-0
> for (i in 1:Nsim) {
+ u1=rnorm(1)
+ u2=rnorm(1)
+ u3=rnorm(1)
+ TTR<-function(a1,a2,a3) {
+ PCE<- 30.2256 -13.5806  *a1 -4.3718  *a2+ 2.4267*a3+
+ 3.9472 *(a1^2-1)-0.9645*(a2^2-1)+4.7846*(a3^2-1)-2.4691
*(a1*a2)+2.2082   *a1*a3-0.8439 *(a2*a3) }
+   t<-TTR(u1,u2,u3)
+   X1[i]<-t  }
>    X1
```

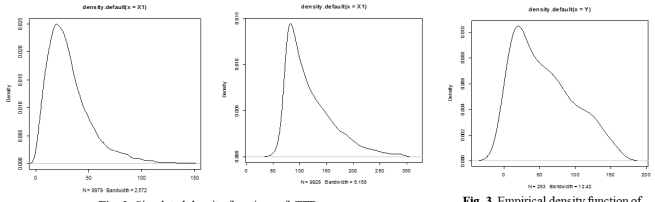


Fig. 2. Simulated density functions of TTR

Fig. 3. Empirical density function of TTR

The survival functions (disease-free survival time) can be estimated by: $S(t) = 1 - F(t)$, where F is the distribution function of TTR (Fig.4 and Fig. 6) (with *Ecdf* function from *Hmisc* package) and this function can be compared with the Cox model (Fig. 5 and Fig. 7).

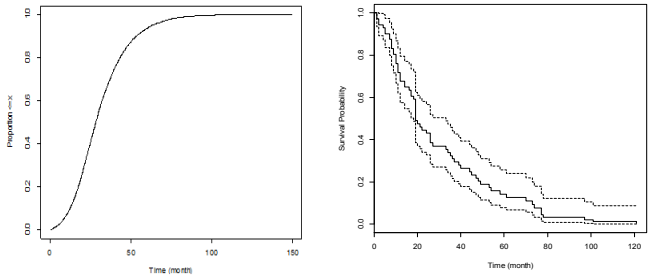


Fig. 4. Simulated survival function only with remission data

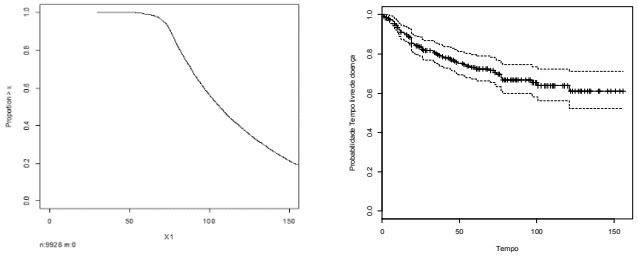


Fig. 6. Simulated survival function with censored data

To estimate the hazard function, which is defined by $H(t) = \frac{f(t)}{S(t)}$, $f(t)$ being the density function of TTR, we need to get the theoretical distribution that best fits the survival functions $S(t)$ (Fig.7 and Fig. 9). These were Gamma and Lognormal functions, respectively. Once identified the distributions that best fit the survival functions, it is possible to estimate H and obtain its graphical representations (Fig. 10 and Fig. 11).

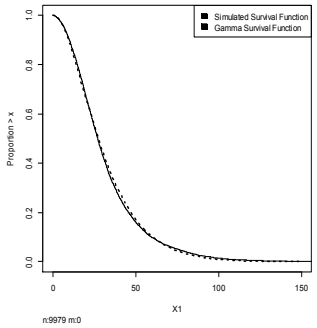


Fig. 8: Survival function adjusted with simulated only with remission data (obtained from the Gamma PDF fitted to simulated data.)

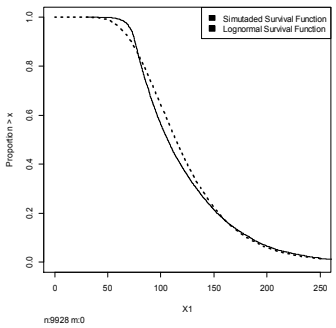


Fig.9: Survival function adjusted with simulated censored data (obtained from the Lognormal PDF fitted to simulated data.)

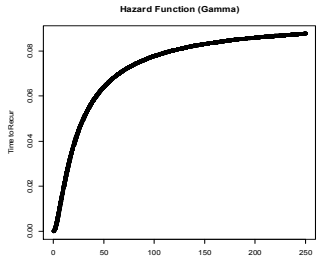


Fig. 10. Hazard function (obtained with Gamma PDF and survival function fitted only with simulated remission data)

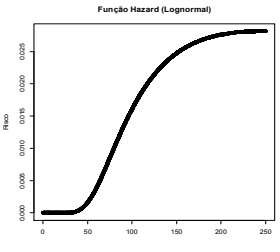


Fig.11. Hazard function (obtained with Lognormal PDF and survival function fitted with simulated censored data)

V. Considerations and Conclusion

The application of SRSM in the computer modelling of complex systems is crucial, since the full model simulation usually requires high computational costs. The methodology has been applied mainly to the analysis of environmental risk problems in particle transport and fluid, and on the analysis of structural reliability. In this work its application to a problem of survival analysis, in connection with morphological features of cell nuclei of breast tumour data, was considered.

A model with Hermite polynomials was fitted to estimate the time to recurrence of breast cancer after tumor excision, according to three characteristics of the malignant cells nuclei: extreme area, extreme texture and tumor size. Two studies were performed: the first one only with recurrence data, and therefore uncensored data, and second, adding to these the censored data, data from patients who had no recurrence until the end of the study or those that left the study before it has finished.

In the first study it was expected that results would suffer from a bias. As indicated by [37], most observed recurrences were at relatively short time (a mean of 24 months) and, therefore, the simulation on the model obtained with a regression method using just these data should result in low prediction of recurrence time, confirming the bias of this particular data set (mean of 31 months). However, it should be noted that [22] refer that it is expected a peaks of recurrence about 18 months after surgery.

In the censored data, the higher probability of recurrence occurs not very far from 60 months and mean of TTR is about 119 months. The peak at 60th month is often referred in the literature, see for example [18] and [17]. In the survival function it can be observed that it is about the sixtieth month that the probability of recurrence begins to decrease significantly, coinciding with that peak of occurrence.

Acknowledgment

This work was partially sponsored by national founds through the Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT under the project (PEst-OE/MAT/UI000 6/2014).

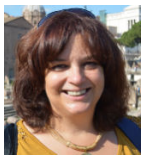
References

[1] A. Ahmed, A. H Soubra, "Extension of subset simulation approach for uncertainty propagation and global sensitivity analysis." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 6.3: pp. 162-176, 2012.

[2] A. Y.Sun, M. Zeidouni, J. P. Nicot, Z. Lu, D. Zhang, "Assessing leakage detectability at geologic CO< sub> 2</sub> sequestration sites using the probabilistic collocation method", Advances in Water Resources, 56, pp. 49-60, 2013.

[3] B. G.Koerkamp, M. C. Weinstein, T. Stijnen, M. H. Heijenbrok-Kal, M. M. Hunink, "Uncertainty and patient heterogeneity in medical decision models", Medical Decision Making, 30(2), pp.194-205, 2010.

[4] C. Leal, T. A. Oliveira and A. Oliveira, "Stochastic Response Surface Methodology: A study on polynomial chaos expansion", SMTDA Proceedings, Lisbon. In Press, 2014.
[5] C. Leal, T. A. Oliveira and A. Oliveira, "Stochastic Response Surface Methodology: A Study in the Human Health Area", in T. E. Simos, G. Psihoyios, Ch. Tsitouras and Z. Anastassi (eds.), Numerical Analysis and Applied Mathematics ICNAAM 2014, AIP Conference Proceedings, American Institute of Physics, to be published, 2014.
[6] D. Datta, H.S. Kushwaha, "Uncertainty Quantification Using Stochastic Response Surface Method Case Study-Transport of Chemical Contaminants through Groundwater", International Journal of Energy, Information & Communications, 2(3), 20011.
[7] D. Datta, "Uncertainty modeling of retention function in biokinetic model using polynomial chaos theory-development of computational algorithm", International Journal of Mathematical Archive (IJMA) ISSN, pp.2229-5046, 4(4), 2013.
[8] D. Li, Y. Chen, , W. Lu, C. Zhou, "Stochastic response surface method for reliability analysis of rock slopes involving correlated non-normal variables", Computers and Geotechnics, 2011, 38(1), pp. 58-68.
[9] D. Q. Li, S. H. Jiang, Y. F. Chen, C. B. Zhou, "Reliability analysis of serviceability performance for an underground cavern using a non-intrusive stochastic method", Environmental Earth Sciences, 71(3), pp. 1169-1182, 2014.
[10] D. Xiu, G. E. Karniadakis, "Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos", Computer Methods in Applied Mechanics and Engineering, 191(43), pp. 4927-4948, 2002a.
[11] D. Xiu, G. E. Karniadakis, "The Wiener--Askey polynomial chaos for stochastic differential equations", SIAM Journal on Scientific Computing, 24(2), 619-644, 2002b.
[12] D. Xiu, G. E. Karniadakis, "Modeling uncertainty in flow simulations via generalized polynomial chaos", Journal of Computational Physics, 187(1), pp 137-167, 2003a.
[13] D. Xiu, G. E. Karniadakis, "A new stochastic approach to transient heat conduction modeling with uncertainty", International Journal of Heat and Mass Transfer, 46(24), pp.4681-4693, 2003b.
[14] E. Bastug, A. Menafoglio, T. Okhulkova, "Polynomial Chaos Expansion for an Efficient Uncertainty and Sensitivity Analysis of Complex Numerical Models", Conference Paper. ESREL 2013, Amsterdam, Netherlands, 2013.
[15] G. Blatman, "Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis", Ph.D Thesis, Université Blaise Pascal, Clermont-Ferrand ,2009.
[16] I. Ardoino, E. M. Biganzoli, C. Bajdik, P.J. Lisboa, P. Boracchi, F. Ambrogi, Flexible parametric modelling of the hazard function in breast cancer studies. Journal of Applied Statistics, 39(7), 1409-1421, 2012.
[17] I. Jatoi, A. Tsimelzon, H. Weiss, G. M. Clark, S. G. Hilsenbeck, "Hazard rates of recurrence following diagnosis of primary breast cancer", Breast cancer research and treatment, 89(2), 173-178, (2005.
[18] M. W. Retsky, R. Demicheli, D. E. Swartzendruber, P. D. Bame, R. H. Wardwell, G. Bonadonna,, ... P. Valagussa, "Computer simulation of a breast cancer metastasis model", Breast cancer research and treatment,45(2), 193-202, 1997.
[19] N. Wiener, "The homogeneous chaos", American Journal of Mathematics, 60(4), pp. 897-936, 1938.
[20] O. G. Ernst, A. Mugler, H. J. Starkloff, E. Ullmann, "On the convergence of generalized polynomial chaos expansions", ESAIM Math. Model. Numer. Anal, 46(2), pp.317-339, 2012.
[21] O. L. Mangasarian, W. N. Street, W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming", Operations Research, 43(4), pp. 570-577, 1995.
[22] R. Demicheli, A. Abbattista, R. Miceli, P.Valagussa, G. Bonadonna, "Time distribution of the recurrence risk for breast cancer patients undergoing mastectomy: further support about the concept of tumor dormancy", Breast cancer research and treatment, 41(2), pp. 177-185,1996.
[23] R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, 2012.
[24] R. G. Ghanem, P. D. Spanos, "Stochastic finite elements: a spectral approach (Vol. 41). New York: Springer-Verlag, 1991.
[25] R. S. Dittus, S.D. Roberts, J. R. Wilson, "Quantifying uncertainty in medical decisions", Journal of the American College of Cardiology, 14(3), A23-A28, 1989.
[26] S.Balakrishnan, A. Roy, M. G. Ierapetritou, G. P. Flach, P.G. Georgopoulos, "Uncertainty reduction and characterization for complex environmental fate and transport models: An empirical Bayesian framework incorporating the stochastic response surface method", Water Resources Research, 39(12) , 2003.
[27] S. Oladyshkin, H. Class, R. Helmig, W. Nowak. "Highly Efficient Tool for Probabilistic Risk Assessment of CCS Joint with Injection Design", Computational Geosciences, 13, 451-467, 2009.
[28] S. Oladyshkin, W. Nowak, "Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion", Reliability Engineering and System Safety,106, pp.179-190, 2010.
[29] S. Oladyshkin, H. Class, R. Helmig, W. Nowak, "A concept for data-driven uncertainty quantification and its application to carbon dioxide storage in geological formations", Advances in Water Resources, 34(11), pp.1508-1518, 2011.
[30] S. Oladyshkin, W. Nowak, "Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion", Reliability Engineering & System Safety, 106, pp.179-190, 2012.
[31] S. Oladyshkin, H. Class, W. Nowak, "Bayesian updating via bootstrap filtering combined with data-driven polynomial chaos expansions: methodology and application to history matching for carbon dioxide storage in geological formations", Computational Geosciences, 1-17, 2013.
[32] S .S. Isukapalli, S. S., Roy, A., Georgopoulos, P. G.: Stochastic response surface methods (SRSMs) for uncertainty propagation: application to environmental and biological systems. Risk analysis, 18(3), pp. 351-363, 1998.
[33] S.S. Isukapalli, "An uncertainty analysis of transport transformation models", Ph.D. Thesis, New Brunswick, New Jersey: The State University of New Jersey,1999.
[34] T.A. Oliveira, C. Leal, A. Oliveira, "Stochastic Response Surface Methodology: A Study in the Human Health Area". ICNAAM Proceedings, Rhodes-Greece. In Press, 2014.
[35] T.A Oliveira, C. Leal and A. Oliveira, " Response Surface Methodology: a review of applications to risk assessment", in Kitsos, C., Oliveira, T., Rigas, A. and Gulati, S. (eds.), Chapter XXIX in Theory and Practice of Risk Assessment, Springer Proceedings in Mathematics and Statistics, to be published (2015).
[36] W. H. Wolberg, W. N. Street, O. L. Mangasarian, "Importance of nuclear morphology in breast cancer prognosis", Clinical Cancer Research, 5(11), pp.3542-3548, 1999.
[37] W. N. Street, O. L., Mangasarian, W. H. Wolberg, "An inductive learning approach to prognostic prediction" In ICML (pp. 522-530) ,1995.
[38] Wisconsin Breast Cancer Prognosis Dataset. http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html#prog



Teresa A. Oliveira

Teresa A. Oliveira is Assistant Professor of Statistics in the Department of Sciences and Technology, at the Universidade Aberta (UAb), Portugal, www.uab.pt, since 2000. She coordinates de Master in Biostatistics and Biometry. She holds a PhD in Statistics and Operations Research - Experimental Statistics and Data Analysis (University of Lisbon-2000), and MS in Statistics and Operations Research (Faculty of Sciences, University of Lisbon-1994). Her research interests include Experimental Design, Statistical Modelling, Data Analysis and Risk Analysis as well as Mathematical and Statistical e-Learning. Co-editor of the volumes "Risk Assessment Challenges: Theory and Practice", by "Springer Proceedings in Mathematics and Statistics" and of the volume "Statistical and Biometrical Challenges: Theory and Applications", she has published several papers in international journals, books and proceedings. She has supervised several PhD and master thesis regarding her interest fields. She co-organized several international conferences on Statistics, Mathematics and Computation (8th Edition - March 2014), as well as Portuguese-Polish Workshops on Biometry (4rd Edition September 2014). She is an integrated member of the Center of Statistics and Applications of the University of Lisbon and she is a collaborator in UIED-Research Unit on Education and Development, Faculty of Sciences and Technology of New University of Lisbon, in IPM- Preventive Medicine

Institute, Faculty of Medicine of University of Lisbon and in LEaD - Laboratory of Distance Learning of Universidade Aberta. She is a member of Portuguese Statistical Society SPE), International Biometric Society (IBS) and International Statistical Institute (ISI). She is the scientific secretary of the Committee on Risk Analysis of the International Statistical Institute (CRA-ISI), see http://www.isi-web.org/44-com/com/126-ra from January 2014.

Conceição Leal

Conceição Leal teaches Mathematics in Escola Secundária de Paços de Ferreira, Portugal, since 1991. She was Department Coordinator and Principal Investigator of the Self-School Evaluation. She holds a degree in Mathematics and a pos-graduation on Statistics, Mathematics and Computation by Universidade Aberta (UAb), Portugal. She is finishing Master thesis on Statistics, Mathematics and Computation and the pos-graduation on Biostatistics and Biometry. Her main research interests include Pure and Applied Mathematics, Experimental Design, Statistical Modelling, Data Analysis and Risk Analysis. She has participated in several international conferences with oral presentations and she has published several papers in Response Surface area.

Amílcar Oliveira

Assistant Professor of Statistics on courses of 1st and 2nd cycle offered by the Department of Sciences and Technology, Universidade Aberta (UAb), Lisbon, Portugal.
He holds a Ph.D. in Mathematics, (Statistical Modelling), Universidade Aberta, an Master degree. in Statistics and Optimization, Faculty of Sciences and Technology, New University of Lisbon.
His research interests include Statistical Modelling. He has published several papers in international journals, books and proceedings regarding these fields. He is integrated member of the Center of Statistics and Applications, University of Lisbon.

Impact of contact surface on accuracy of humidity distribution measurements in autoclaved aerated concrete constructions by EIS

Sanita Rubene, Martins Vilnitis, Juris Noviks

***Abstract*—**Humidity distribution throughout the cross section of autoclaved aerated concrete masonry constructions has significant impact on its performance of heat resistivity properties. An application of electrical impedance spectrometry (EIS) method for determination of humidity distribution throughout the cross section of autoclaved aerated concrete constructions has been a subject of research recently. The EIS method seems to be a useful and convenient method for detection of the humidity distribution throughout the cross section of a construction. Impact of contact surface between the measurement probe and the testing concrete sample is a subject of this research.

***Keywords*—**EIS, non-destructive testing, humidity distribution, aerated concrete.

I. Introduction

ENERGY efficiency is one of the most actual and significant topics of research. As autoclaved aerated concrete masonry blocks are a construction material that combines load bearing properties as well as high heat insulation properties it is very suitable for use as a load bearing construction material in building where high heat insulation performance is required. The most common problem in use of autoclaved aerated concrete constructions is neglected drying period of freshly manufactured material and lack of useful non-destructive test methods for easy detection of humidity distribution throughout the cross section of the masonry construction.

EIS is a relatively new method for non-destructive measurements of humidity distribution in aerated concrete constructions but it has already displayed acceptable results in field of relative measurements for humidity distribution. Therefore it allows to control the drying process of the masonry construction in order to reach as high heat insulation parameters of the material as it is possible. As the measurements are performed with a pair of probes which are inserted in the construction then the question of impact of contact surface between the measurement probe and the testing concrete sample arises. This is a subject of the research described in this paper.

Sanita Rubene is with the Riga Technical university, Faculty of Civil engineering, Construction Technology department, Riga, LV-1048, Latvia (phone+371 26461876; e-mail: sanita.rubene@inbox.lv).

Martins Vilnitis is with the Riga Technical university, Faculty of Civil engineering, Construction Technology department, Riga, LV-1048, Latvia (e-mail: martins.vilnitis@rtu.lv).

Juris Noviks is with the Riga Technical university, Faculty of Civil engineering, Construction Technology department, Riga, LV-1048, Latvia (e-mail: praktiska@321.lv).

II. Methods

For particular research electrical impedance spectrometry method using Z-meter III device was used.

Method of electrical impedance spectrometry (EIS) enables detection of the distribution of impedance or other electrical variables (such as resistivity, conductivity etc.) inside a monitored object, and thus the observation of its inner structure and its changes [1-2]. This method ranks among indirect electrical methods and it is used in measuring properties of organic and inorganic substances. So far, EIS is widely used in medicine as one of the most common testing methods in diagnostics where any kind of tissues are involved. It constitutes a very sensitive tool for monitoring phenomena that take place in objects (e.g. changes occurring in earth filled dams when loaded by water, in wet masonry sediments etc.), electrokinetic phenomena at boundaries (e.g. electrode/soil grain, between soil grains) or for describing basic ideas about the structure of an inter phase boundary (e.g. electrode/water) [3]. In particular case the measurement results were obtained as a value of real part and imaginary part of electrical impedance. It made easier to determine which changes of impedance were caused by the changes of humidity ratio and which were caused by the changes of materials' porous structure during its drying process. For further measurements the real part of electrical impedance is used.

The EIS is based on the periodic driving signal – the alternating signal. If low amplitude of the alternating signal is used, concentration changes of charge are minimal at the surface of an electrode connected with the measured surface, which is very important in systems sensitive to so called concentration polarization. The range of frequencies used for

ISBN: 978-1-61804-253-8

the driving signal enables the characterisation of systems comprising more interconnected processes with different kinetics.

In the Laboratory of Water – Management Research of the Institute of Water Structures at the Civil Engineering Faculty of Brno University of Technology, a measuring instrument with a Z-meter III device has been developed within the solution of an international project E!4981 of programme EUREKA. This instrument is verified in laboratory experiments and measurements on objects in situ [4-5].

The initial application of the device was for determination of humidity distribution changes and monitoring of moisture migration in earth-filled dams and other similar constructions [6-8].

This far application of the EIS method is credible for measurements of relative changes in humidity distribution throughout the cross section of the autoclaved aerated concrete constructions and further researches are performed in order to develop the method for wide use of non-destructive humidity distribution measurements.

The measurements are performed by inserting a pair of measurement probes in bores made in the construction. The measurements are taken in area between the measurement probes. In the particular case the measurements are performed using 4 (four) or 5 (five) measurement channels (Fig.1) of the probes depending on thickness of the construction. The channels are counted starting from the furthest channel of the probe.

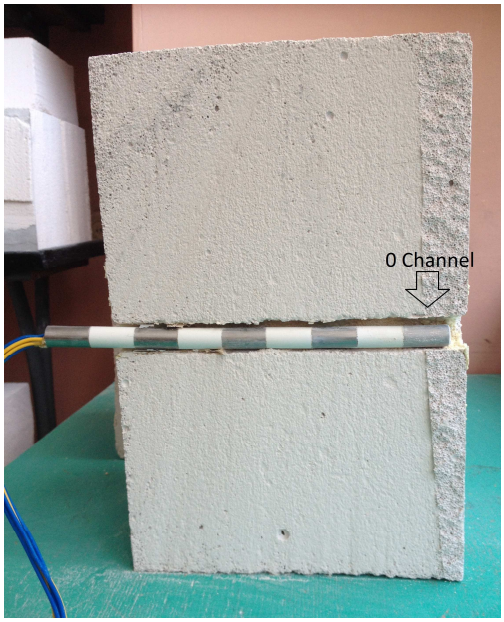


Fig.1. Cross section of autoclaved aerated concrete block used for the experiment with measurement probe

The measurements are performed between the stainless steel elements in both probes which are inserted in the construction within a distance (Fig.2).

III. PREVIOUS RESEARCH IN FIELD OF APPLICATION OF EIS FOR HUMIDITY DISTRIBUTION MEASUREMENTS IN AERATED CONCRETE CONSTRUCTIONS

A series of research have been performed to determine the possibilities to apply EIS method and particularly Z-meter III device for humidity distribution measurements throughout the cross section of autoclaved aerated concrete constructions.

The first research about the possibilities of application of EIS method for humidity distribution measurements in autoclaved aerated concrete constructions by Z-meter III device was performed by authors in 2013 in Brno university of Technology [5]. The tests were performed with different types of measurement probes and proved that humidity distribution changes throughout the cross section of autoclaved aerated concrete masonry block sample correlate with changes of its electrical resistivity measurements. All of the tests were performed as surveillance of changes in resistivity measurements while one side of the block was exposed to direct impact of water. The experiment proved that the EIS method was applicable for detection of humidity distribution changes in relative means.

The second part of the research about the application of EIS method on humidity distribution changes was performed by authors in Riga Technical university [9]. The subject of this research was to ascertain that the results obtained in the experiments performed in the Brno university of Technology [5] were credible as well as to determine the impact of cracks or masonry joints on the EIS measurement results.

Series of tests performed during the previous researches proved that the results of EIS measurement tests in field of humidity distribution throughout the cross section of the masonry block are credible and confirmed the previous results – by the increasing of the humidity ratio in the area of the monitored cross section its electrical resistivity values decreased. As for the second part of the research, it showed that large cracks or masonry joints between the measurement points have significant impact on measurement results in absolute values. It means that measurements performed in masonry construction area without significant cracks or masonry joints cannot be compared with the results that are obtained in areas where the masonry joints or significant cracks are present.

As all previous researches [5,9] were performed on single autoclaved aerated concrete blocks or combinations of two such blocks then the next step of the research was to monitor the drying process of an aerated concrete block masonry wall segment [10]. The results of the research displayed a humidity migration process throughout the cross section of the wall segment and its correlation to the EIS measurements. The obtained data allow visualizing the speed and character of the humidity migration process in the wall segment after its construction until the moment when the air-dry state of the construction is reached in the laboratorial environment. The test approved authors' assumption that the aerated concrete

ISBN: 978-1-61804-253-8

drying properties are efficient in case of the construction which is not covered with finishing layers.

In additional research the impact of masonry joints on the values of the EIS measurements were performed [11-12]. The research allows assuming that there exists linear correlation between the results obtained in the areas where are no masonry joints between measurement points and the results, which are obtained in areas, including masonry joints between the measurement points. The determined correlation indicates that the impact of joints can be described with a linear equation (1) where a quotient is in range of 0,67 to 0,96 but the value of C is depending on the width and filling material (mortar, special glue etc.) of the masonry joint.

$$y = ax + C \quad (1)$$

IV. Experiments

For this research autoclaved aerated concrete masonry blocks with density of 375kg/m^3 were used (Fig.2). [13]

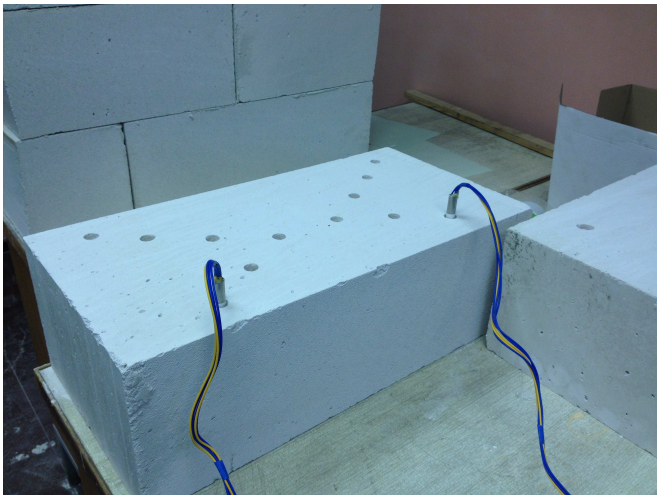


Fig.2. Autoclaved aerated concrete block with one pair of measurement probes

The experiment included usage of three measurement series where in each set different type of contact surface improvements were performed. To increase the credibility of obtained results each of the series consist of measurements performed on three blocks which are exposed to identical conditions.

For the first set of the blocks the measurement bore was made with the maximal accuracy for the bore diameter to comply with the diameter of the measurement probe and it was maximally close to the probe diameter. The difference between the probe radius and the radius of the bore was approx. 1mm (Fig.3).

Fig.3. Autoclaved aerated concrete block used for the experiment

Although such precision causes certain inconveniences in measurement process due to difficult insertion of the measurement probe the contact surface in this case was very tight (Fig.3, 4).

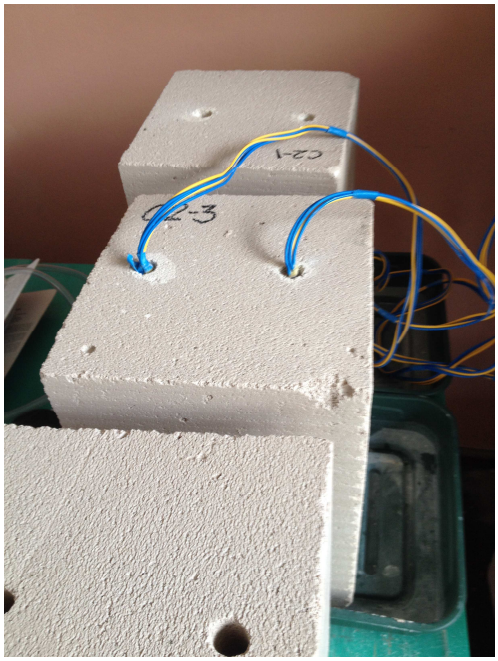


Fig.4. Autoclaved aerated concrete blocks with liquid silicone filling in bores and bores with no filling

In measurement process arises a problem of extensive drying process due to additional ventilation of the blocks through the measurement bores and a solution for this problem was researched during this experiment as well as the problem of contact surface's influence on the precision of measurement results.

In the second set of experiment in order to prevent fast drying of the block due to additional ventilation of the structure in the measurement bores the bore surface was covered with liquid silicone and the influence of this material on measurement results was also assessed.

ISBN: 978-1-61804-253-8

In the third set of the experiment blocks the space in the measurement bore between the probe and the surface of the aerated concrete block was filled with universal silicone. The silicone fills all free space between the measurement probe and the aerated concrete so that ideal contact surface between the probe and the block is established. Although a problem arises during the drying process of the block. As the block dries the small particles of aerated concrete which were displaced in the pores of the material during the drilling process separates from the block and in such way decrease adhesion between the silicone and the block. As the measurement process assumes cyclic insertion and withdrawal of the measurement probe then the decreased adhesion between the materials cause separation of the silicone from the block surface and in such way the integrity of the measurement process is terminated.

The obtained results were compared within each of the measurement sets and afterwards the results were compared between the sets.

All blocks were monitored for period of three months during summer of 2014. The average temperature of the surrounding environment during the monitoring process was +22°C and average air humidity rate was 70%.

V. RESULTS

A. *Impact of the bore diameter on the measurement results*

For all following results the measurement result which was obtained from the EIS measurements from autoclaved aerated concrete blocks with no filling between the block surface and measurement probe is considered as a reference result.

The reference chart is displayed in Fig.5.

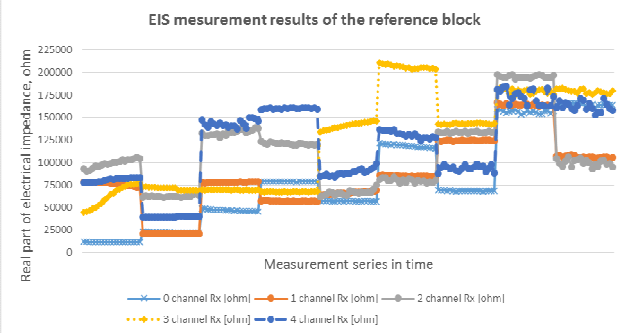


Fig.5 EIS measurements during the drying process of the reference block

During the blocks` drying process its` electrical resistivity changes. Fig.5 displays that the changes of the electrical resistivity are not equal in all areas of the block and tendencies of humidity migration throughout the cross section of the block can be monitored through the EIS measurement results. From the measurement results it can be concluded that each sector of the block show different speed of drying and moisture transfer. The resistivity measurements of separate channels can change significantly in time from high values (aka relatively low humidity rate) to low values (aka relatively high humidity rate) but the tendency of overall increase of the resistivity can be correlated to overall decrease of the materials' relative humidity rate.

B. *Impact of the silicone filling on the measurement results*

As the application of universal silicone significantly improves contact surface between the measurement probe and the surface of aerated concrete block this material was used in the second set of measurements (Fig.6).

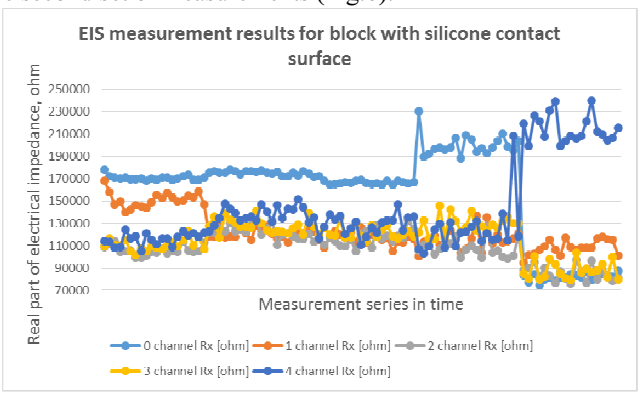


Fig.6 EIS measurements during the drying process of the block with silicone filling between the measurement probe and the surface of the aerated concrete block

After comparing results of Fig.5 and Fig.6 it can be stated that the results for the measurements where silicone was used as filling material of space between the probe and the block surface the overall trend of the charts is similar but the later measurements are more even and do not show significant differences between the measurement series which are taken in different time. In case of not using silicone the differences between the measurement series are obvious (Fig.5).

Comparison of absolute measurement values between the measurements made without any filling material and using silicone as filling show that the silicone filling works as insulation material which does not allow to monitor detailed changes of moisture migration through the cross section of the material in real time (Fig.7).

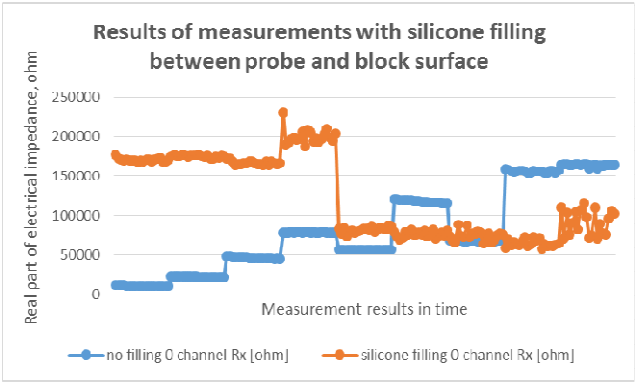


Fig.7 EIS measurements during the drying process of the reference block and block with silicone filling between the probe and the measurement surface

These results also show significant differences in absolute values of measurement data which does not allow to correlate data obtained during different measurement sessions. In fact the measurement results prove that during the drying process

ISBN: 978-1-61804-253-8

silicone layer loses its' adhesion to the measurement surface and then a leap of measurement values can be observed. This brings to a conclusion that due to difficult application of silicone in measurement bores and to the fact that silicone affects measurement results it is not the most suitable material for prevention of measurement surfaces' accelerated drying process or for improvement of the contact surface between the measurement surface and the probe.

C. Impact of the liquid silicone on the measurement results

The experiment with application of universal silicone in order to prevent accelerated drying processed of the measurement areas due to additional ventilation and exposed drying surfaces showed that universal silicone is not suitable for such purpose.

As an alternative a liquid silicone spray was chosen. After the preparation and cleaning of the measurement bores with compressed air a layer of liquid silicone spray was applied on all surface of the measurement bores. The liquid silicone infiltrated into the aerated concrete for approx.1cm (Fig.4, the block with the measurement probes inserted). After the blocks were infiltrated with water the area which was exposed to liquid silicone spray became clearly visible and in such way it can be stated that liquid silicone had covered all area of contact surface between the measurement probe and the aerated concrete material.

Although liquid silicone spray infiltrates into the aerated concrete and thus does not improve the contact surface, it prevents the measurement bore surface from accelerated drying. In such way, these results can provide more precise data about moisture migration processes throughout the cross section of the construction.

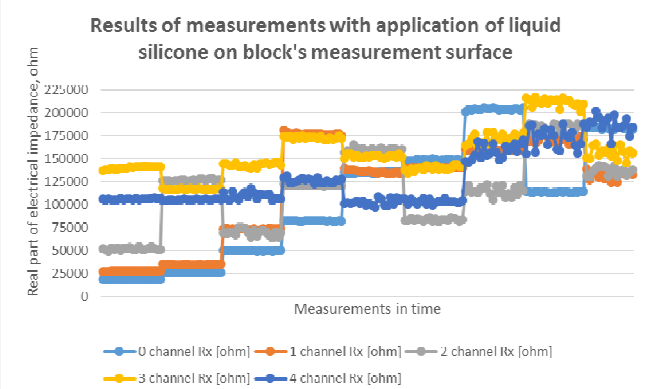


Fig.8 EIS measurements during the drying process of block with liquid silicone spray filling between the probe and the measurement surface

The overall character of the measurement results for the reference block and for the block with liquid silicone application of measurement contact surface is similar (Fig.5 and Fig.8). Exact division into measurement series can be seen in the chart which allow to assume that liquid silicone does not have significant impact on sensitivity of measurement tool as universal silicones does.

Comparing the absolute values of the measurement results (Fig.9) from the reference block and the block with liquid silicone filling on measurement bores it can be stated that liquid silicone spray does not have significant impact on absolute values of EIS measurement results.

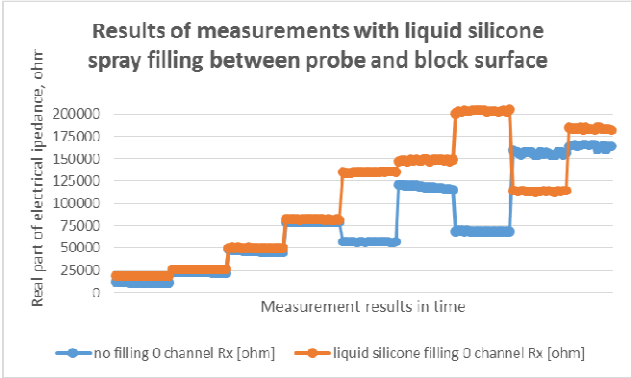


Fig.9 EIS measurements during the drying process of the reference block and block with liquid silicone spray filling between the probe and the measurement surface

Furthermore, it can be stated that application of liquid silicone spray on measurement bore surface excludes some influence of surrounding area to the results of the experiment. The barrier which is made by the silicone covering prevents measurement surface from accelerated drying in warm and dry conditions and increased absorption of air humidity in case of high humidity rate of surrounding environment.

VI. CONCLUSION

Electrical impedance spectrometry is a convenient method for non-destructive determination of humidity distribution throughout the cross section on autoclaved aerated masonry constructions. This research proved that the contact surface between measurement probe and the measurement surface of the masonry block has significant influence on measurement results. If the filling material for improvement of the contact surface is chosen wrongfully then it can lead to misinterpretation of the measurement data or even for obtaining of wrongful data. The example with universal silicone proved that there exist such materials which can reduce accuracy of the obtained results and contrary to expected result that it will improve the accuracy of obtained data it had an opposite effect.

Experiment proved that from the three types of contact surface measurements were performed on the most precise results can be obtained from the contact surface on which liquid silicone spray is applied. The liquid silicone penetrates into the porous structure of autoclaved aerated concrete material for approx.1cm and establishes a layer within the contact surface that prevents excessive drying of the measurement area though the surface of measurement bores and in separate cases prevents recurrent absorption of humidity from the surrounding environment. The measurement values on these measurements almost do not differ from the measurement values of the reference block. It means that in

ISBN: 978-1-61804-253-8

cases when it is important to establish correlation between the electrical impedance spectrometry measurements and absolute values of relative humidity rate in the material standard correlations can be applied for the interpretation of the results. It is important for correct interpretation of the obtained results because standard correlations are prepared with blocks where no contact surface improvement is used.

ACKNOWLEDGMENT

This research was performed in Riga Technical university, Faculty of Civil engineering. The "Z-meter III" device was invented in framework of EUREKA E!4981.

REFERENCES

[1] McCarter, W.J.; Garvin, S. Dependence of Electrical Impedance of Cement-Based Materials on their Moisture Condition. In: Journal of Applied Physics Series D: Applied Physics 22 (1989), No. 11, S. 1773-1776.

[2] Elsener, B. Ion Migration and Electric Conductivity in Concrete. Zürich : Schweizerischer Ingenieur- und Architekten-Verein, 1990. - In: Korrosion und Korrosionsschutz. Tl 5. Electrochemical Protection Process for Concrete Building Structures, Symposium 15. November 1990, S. 51-59.

[3] Skramlik, J., Novotny, M.. One-dimensional moisture transport monitored by a non-destructive method. INTERNATIONAL JOURNAL OF COMPUTERS Issue 4, Volume 2, 2008.

[4] Parilkova, J. et al. Monitoring of changes in moisture content of the masonry due to microwave radiation using the EIS method. EUREKA 2011, ISBN 978-80-214-4325-9, Brno (2011).

[5] Rubene S. et al. DETERMINATION OF HUMIDITY LEVEL IN AERATED CONCRETE CONSTRUCTIONS BY NON DESTRUCTIVE TESTING METHODS, in proc "Innovative materials, structures and technologies" Riga, 2014 p.135-140.

[6] Parilkova, J. The EIS Method and a Z-meter III Device, a lecture within an event in Litice.;

[7] Parilkova, J., Pavlik, J. An automated system for analysis of selected characteristics and processes in a porous environment using the EIS method. A partial report of the Project OE10002 for the year 2010 externally examined, Brno (2010);

[8] Parilkova J. et al. Optimization of methods for monitoring the unconfined water table and its action in earth-fill dams. Sub – report of a project of the Grant Agency of the CR 103/04/0741, LVV UVST FAST VUT v Brne (2005).

[9] Rubene S., Vilnitis M. APPLICATION OF ELECTRICAL IMPEDANCE SPECTROMETRY FOR DETERMINATION OF MOISTURE DISTRIBUTION IN AERATED CONCRETE CONSTRUCTIONS in proc. "1st conference and working session EUREKA!7614" Brno, Czech Republic, 2013. p. 124-130.

[10] Rubene, S., Vilnītis, M., Noviks, J. Monitoring of the Aerated Concrete Construction Drying Process by Electrical Impedance Spectrometry. In: Proceedings of 4th International Conference "Advanced Construction 2014", Lithuania, Kaunas, 9-10 October, 2014. Kaunas: Kaunas University of Technology, 2014, pp.216-220. ISSN 2029-1213.

[11] Rubene S., Vilnitis M. MONITORING OF HUMIDITY DISTRIBUTION CHANGES IN AERATED CONCRETE MASONRY CONSTRUCTION BY EIS in proc. "2nd conference and working session EUREKA!7614" Brno, Czech Republic, 2014.

[12] Rubene, S., Vilnītis, M., Noviks, J. Impact of masonry joints on detection of humidity distribution in aerated concrete masonry constructions by electric impedance spectrometry measurements. Submitted for publication in: Proceedings of ICCEAUE 2014 : XII International Conference on Civil , Environmental, Architectural and Urban Engineering, UK, London, 2014.

[13] http://aeroc.lv/index.php?page=938&lang=lat&cnt=AEROC_Universal (site accessed on October 28, 2014).

Quantum Vacuum Dynamics, Coherence, Superluminal Photons and Hypercomputation in Brain Microtubules

Luigi Maxmilian Caligiuri and Takaaki Musha

***Abstract*—**Recent researches have shown the theoretical possibility of accomplishing hypercomputation in human brain by using eventual superluminal evanescent photons generated inside brain's microtubules, using these as quantum waveguides or resonant cavities. Nevertheless, no convincing physical mechanisms has been proposed so far, able to explain the generation of such eventual superluminal photons inside microtubules and the possibility to eventually use them to manipulate quantum bits in brain. In this paper we propose a novel theoretical model according to which a confined field of faster than light photons field of suitable wavelength can arises from a spontaneous phase transition of the QED quantum vacuum occurring in the water contained inside the microtubules inner volume. It has been shown that, in the water trapped inside microtubules, there exist the conditions for the formation of a macroscopic coherent quantum state in which water molecules oscillate in phase with an e,m, field associated to a suitable electronic transition. We have also discussed some interesting consequences of these results on the possibility of hypercomputing in human brain.

***Keywords*—**Coherence, Decoherence, Evanescent Photons, Hypercomputing, Microtubules, Phase Transition, Quantum Vacuum, Superluminal Photons, Superradiance, Water.

I. Introduction

THE connection between brain's functions and quantum physics has become more and more evident due to several results from the latest research developments.

For many time the possibility to interpreter, in the light of quantum physics (QP), such functions (as well as many other physical phenomena) was hindered by the deeply erroneous conviction that quantum physics was limited to the description of microscopic world. The fallacy of this belief firstly lies in the ascertainment that QP actually concerns not just phenomena occurring in the microscopic world only but, more correctly, all the phenomena involving quantized exchanges of energy. Fundamental examples coming from condensed matter physics [1,2,3,4] have proved the rightness of this statement as, for example, the discovery of superconductivity (cold and hot) and ferromagnetism, all showing the presence of a quantum behavior over macroscopic (with respect the atomic or molecular size) distances and at relatively high temperatures.

One of the most fascinating connection between QM and biological processes could be related to the emergence of consciousness, the formation of memory and computational functions in human brain. In particular, the unity and non-local manifestations of consciousness [5-7] could be naturally explained in terms of quantum coherence and long-range correlation as well as the non-algorithmic and non-computational features of certain brain functions, as argued by Penrose [8,9]. Quantum mechanics is also a good candidate to explain the unique distributed features of some brain's functions as memory storage and perceptual processing for example in connection to the so-called holographic model of brain firstly proposed by Pribram [10] in which such functions are linked to the associative features of parallel distributed processes at the basis of holographic optical techniques. The general requirement of non - locality and cooperative activity is also strongly suggested by the persistence of memory and perception functions in brain, even in the presence of extensive tissue damage [11].

After the Schrodinger's intuition, expressed in his book "What is life?" [12], the first systematic approach to a quantum interpretation of biological process dated back to the model proposed by Frohlich [13] several years ago. He suggested the occurrence, in biological matter, of macroscopic quantum phenomena able to explain the energy transport without loss in the living organisms and signal transfer based on collective coherent oscillations associated to a branch of longitudinal electric modes in the frequency range

$$\nu_{coh,F} \sim 10^{11} - 10^{12}\, s \tag{1}$$

also known as Frohlich frequency. So in living systems, according to Frohlich's model, energy could be stored in a thin two - dimensional layer placed beneath the cell membrane, that in this way would act as a biological superconducting medium, under dipolar propagating waves without thermal loss [13].

Subsequently Popp [14], basing on a different standpoint, proposed the idea that also the weak emission of photons in the visible range of electromagnetic spectrum by living organisms, the so-called "biophotons", could be related to a sort of coherent mechanism typical of living systems.

Luigi Maxmilian Caligiuri is with Foundation of Physics Research Center (FoPRC), Celico, CS 87053 ITALY and University of Calabria, Arcavacata di Rende, CS 87036 ITALY (phone: +39094431875; fax: +390984431875; (e-mails: caligiuri@ foprc.org; max.caligiuri@gmail.com).

Takaaki Musha is with Advanced Science – Technology Research Organization, Kanagawa-ken, Yokohama, JAPAN and with Foundation of Physics Research Center (FoPRC), Celico, CS 87053 ITALY.(e-mail: takaaki.mushya@gmail.com).

The consistent interpretations of quantum behavior in living organisms given by Frohlich's and Popp's models also found a further theoretical foundation in the analysis given by Umezawa [15] within the framework of quantum field theory, who suggested the presence, in brain's cells, of a spatially distributed system characterized by the full range of quantum mechanical degrees of freedom subjected to quantum phenomena. Later Davydov [16] extended some considerations of the Frohlich's model, whose dipolar coherent oscillations were restricted to thin layers adjacent to cell membranes, by proposing the existence of "solitonic excitation states" responsible for the dissipation-free energy waves propagating along α - helices in proteins.

In 1990's Hameroff and Penrose [8,9,17] suggested a primary role of quantum effects in brain functioning and, in particular, in the emergence of consciousness, by considering the dynamics of microtubules (MT) of brain cells. In their model they pictured the tubulin dimer units of MT as quantum systems described by a coherent superposition of two-levels quantum states, corresponding to the two conformations of tubulin proteins (α - and β -tubulin), depending on the position of the unpaired charge of $18\,e$ relative to pocket.

In particular, Hameroff suggested [18] that MT could be considered as waveguides for photons and as holographic information processors, also due to their periodic lattice structure (providing "periodically arrayed slits") through which photons can pass. From the standpoint of QP, this model [15] considers the phenomenon of the so-called "superradiance" potentially occurring, under suitable conditions, in the water molecules contained in the MT inner volume, namely a quantum collective behavior of water molecules and electromagnetic field modes able to convert the perturbative thermal and molecular disordered oscillations into coherent photon modes inside MT, whose first theoretical treatment was due to Del Giudice, Preparata and Vitiello [19] by the model of water as a free electric dipole laser.

According to the latter, in fact, the coherent interaction between the "matter field" associated to water molecules and a self-generating quantized electromagnetic field arising from QED quantum vacuum, spontaneously occurring under suitable boundary condition on matter density and temperature, is very strong within a spatial region of the order of

$$L \sim 2\pi/\omega_0 \tag{2}$$

where ω_0 (in the units system in which $c = \hbar = 1$) is the energy associated to the transition between a given couple of levels of energy of the matter quantum field corresponding, for the transitions between the ground state and the low lying states of water molecule rotational energy spectrum (of the order of $4\,meV$) , to few hundreds of microns. Within this spatial region, also know as "coherence domain" (CD), the time scale associated to coherent interaction is of order of $10^{-14}\,s$, namely much shorter than the typical time scale connected to short-range interactions.

This coherent dynamics also determines an extended oscillation polarization field able to correlate an high number of water electric dipoles. In this way the coherent interaction between water electric dipoles and the radiating electromagnetic field is able to generate stable and ordered structures in macroscopic spatial regions. Furthermore, due to coherence, the photons associated to this electromagnetic field would be characterized by the so-called "self-induced transparency" according to which they are able to penetrate the optical medium where they propagate as it were made transparent by the photon field itself thus potentially leading, inside cytoskeletal MT of brain, to a sort of coherent optical supercomputers able to enormous elaboration capabilities.

In particular, according to Jibu et al [11], the photons composing the coherent electromagnetic field inside CDs entangle the cytoskeletal protein and the MT quantum sates of a given neuron link to those of other neurons by the tunneling of such coherent photons through biological membranes. In this way these authors argued that also consciousness could arise from such coherent dynamics and, in particular, from the creation-annihilation dynamics of a finite number of evanescent photons in brain.

As pointed out by Recami [20,21] tunneling photons moving in an evanescent field can be characterized by a superluminal group velocity or, equivalently by a negative square mass of the photons belonging to the evanescent field that can be shown by considering that a quantum evanescent photon satisfies the Klein-Fock-Gordon equation, namely (in one dimension):

$$\left[-\left(1/c^2\right)\left(\partial^2/\partial t^2\right) + \nabla^2 - m_0^2 c^2/\hbar^2\right]\psi(x,t) = 0 \tag{3}$$

where c is the velocity of light in vacuum, m_0 the absolute value of the proper mass of the evanescent photon. The solution of (3) is given by

$$\psi(x,t) = A \cdot \exp\left[-\left(p \cdot x + E \cdot t\right)/\hbar\right] \tag{4}$$

corresponding to a particle characterized by an imaginary rest mass $i \cdot m_0$ moving at a superluminal velocity and satisfying the relativistic relation

$$E^2 = p^2 c^2 - m_0^2 c^4 \tag{5}$$

where, as usual, E is the total energy and p the momentum of the particle.

Then, according to this picture, superluminal photons traveling in an evanescent mode arising from coherent macroscopic quantum system inside MT could be able to realize, through a long-range order in living systems, an optical computing network in which brain's MTs can achieve quantum bit computation on large data set, so practically realizing hypercomputing performance with respect conventional processors as suggested by Musha [22]. Furthermore, as shown by Caligiuri [23] in the general context of Special Theory of Relativity and by Ziolkowski [24] in relation to electromagnetic metamaterials, superluminal propagation in principle doesn't violate causality.

Nevertheless Georgiev [25] observed the wavelength of coherent photons associated to the energy difference of $4\,meV$ between the two eigenstates of water molecule

involved in superradiant water model, equal to $\lambda \simeq 310\,\mu m$ (i.e. lying in the infrared range of electromagnetic spectrum), is not comparable with the typical length $l \sim 1\,\mu m$ of a moderate sized MT, being $\lambda \gg l$.

Consequently, according to this last statement, there wouldn't be any nodes and anti-nodes inside MT and there could be no way to use superradiant emission of infrared photons to manipulate the qubits inside the MT cavities or centrioles in a fashion similar to the use of standing wave lasers in the ion trap computation.

In addition, Mavromatos [26], also referring to the Hameroff-Penrose model, recently discussed the role of environmental decoherence on the quantum system composed by brain's MT, observing that, even for in vivo MT, the effect of environmental interaction on this coherent system cannot be ignored and proposed a "quantum electrodynamics cavity model" for MT, based on the consideration of the electromagnetic interaction, at a frequency $\omega \sim 6 \cdot 10^{12}\,Hz$, between the electric dipole moments of tubuline protein dimer units and the corresponding dipole quanta in the thermally isolated water inside the brain's MT, for which he calculated an environmental decoherence time of $O\left(10^{-6} - 10^{-7}\right)s$, a time scale much shorter than that required for conscious perception, but sufficient to allow a loss-free energy transfer and signal propagation along a moderately long MT ($l \sim 1\,\mu m$).

In this paper we propose an alternative theoretical model, already successfully applied by the author to the study of biophotons emission [27], also based on the consideration of the QED coherence in the water inside brain's MT, but considering electronic transitions rather than rotational energy levels in water molecules, able to overcome both the critical points revealed by Georgiev and Mavromatos as well as to explain the origin of superluminal evanescent photons capable to be used for the manipulation of qubits in brain.

The results so far obtained confirm the idea that coherent dynamics of water inside MT could play a primary role in the establishment of long-range order in living organisms and in the formation of high-grade functions in brain as, for example, hypercomputation and, eventually, consciousness.

II. A SYNTHETIC OVERVIEW OF QED COHERENCE IN CONDENSED MATTER

A. Quantum Vacuum fluctuations and energy shift in atoms and molecules

It is a well-known phenomenon in modern physics that the energy of a Hydrogen atom varies as a consequence of the coupling of the electric current associated to the orbiting electron to the electromagnetic field of the QV fluctuations. This effect. called "Lamb – shift", discovered in 1945 and only later understood, demonstrates, together with other experimental evidences (as, for example, the Casimir effect and the radiative correction of the particles masses) the direct interaction between QV and atoms and that this interaction is able to modify the energy of the latter even meaningfully. In modern physics, in fact, the physical vacuum cannot be considered, due to Heisenberg uncertainty principle, as a void but as a physical entity manifesting a complex and fundamental background activity in which, even in the absence of matter, processes like virtual particle pair creation – annihilation and electromagnetic fields fluctuations, known as Zero Point Field (ZPF) or QV fluctuations, continuously occur.

According to the framework of QED coherence in condensed matter, originally developed by Preparata [28] and applied to living systems by Preparata, Del Giudice et al. [29-37], starting from a well-known behavior of electromagnetic and matter quantum fields, under suitable boundary conditions (almost always verified in the condensed matter and living organisms as well), a coherent electromagnetic field, oscillating in tune with all the matter constituents, spontaneously emerges from the self-produced electromagnetic field.

In particular it has been shown [28] that, above a critical density $\left(N/V\right)_{crit}$ and below a threshold temperature T_0, an ensemble of atoms or molecules, placed in the empty space (namely without any matter or radiation field different than ZPF), spontaneous "decays" into a more stable state (characterized by lower energy and so strongly favored) in which all the matter components are phase correlated among them by means of the action of an electromagnetic field oscillating in tune with them too, confined within a defined spatial region, called "Coherence Domain" (CD), associated to the wavelength of the tuning electromagnetic field.

The arising of this physical coherent state can be understood by considering that, according to quantum field theory, matter and fields continuously perform quantum fluctuations. The same types of fluctuations also characterize, as seen above, the QED QV.

We consider the matter system to be composed of electrical charged particles (electrons and nuclei) characterized by a discrete energy spectrum $\left\{E_i\right\}$ and indicate with "0" its fundamental state (whose energy is $E_0 = \hbar\omega_0$) and with "k" a generic excited state (with an associated energy $E_k = \hbar\omega_k$). A vacuum fluctuation able to coupled to the systems and excite the state k (from the fundamental one) must then have a wavelength $\lambda = hc/\delta E$ where $\delta E = E_k - E_0$. The probability of this coupling with the excitation of state k is quantified by the "oscillator's strength" for the transition $0 \to k$, given by [28]

$$f_{0k} = \left(2m_e/3\omega\left|E_k - E_0\right|\right)\sum_j \left|\left\langle 0\left|\vec{J}_j\right|k\right\rangle\right| \qquad (6)$$

where ω is the frequency of the exciting electromagnetic field, m_e the electron mass and $\vec{J}$ the electromagnetic current density operator. For an atom or molecule with n electrons, f must follow the rule

$$\sum_k f_{0k} = n \tag{7}$$

Now let's consider the volume of space $V = \lambda^3$ "covered" by an oscillation of the QV electromagnetic field resonating with them, supposing it contains N atomic or molecular species, and let be P the "Lamb – shift type" probability that a photon "escapes" from QV, couples with an atom or molecule and puts it in a given excited state. The overall probability of coupling for the N constituents is then

$$P_{tot} = P \cdot N = P\left(N/V\right)V = P\left(N/V\right)\lambda^3 \tag{8}$$

that is proportional to the matter density. So, when density exceeds a particularly high value, almost every ZPF fluctuation couple with the atoms or molecules in the ensemble. This condition starts the "runaway" of the system from the perturbative ground state, in which matter and quantum fluctuations are uncoupled and no tuning electromagnetic field exists, to a coherent state in which, within a CD, a coherent electromagnetic field oscillated in phase with matter determining a macroscopic quantum state in which atoms and molecules lose its individuality to become part of a whole electromagnetic field + matter entangled system.

B. The equations of coherence and the "runaway" towards the CGS

The evolution of such a system can be characterized mathematically [28,30] considering, for simplicity, a two-levels matter system described by the matter field $\chi_l\left(\vec{x},t\right)$ with $l = 0,k$ and an electromagnetic field characterized by its vector potential $\vec{A}\left(\vec{x},t\right)$. If we neglect the spatial dependence of both the fields (since they can be assumed slowly varying within the CD) the dynamic equations, describing the time-evolution of the electromagnetic field+ matter interacting ensemble, are given by

$$\begin{gathered} i\dot{\chi}_0\left(\tau\right) = g\chi_k\left(\tau\right)A^*\left(\tau\right) \\ i\dot{\chi}_k\left(\tau\right) = g\chi_0\left(\tau\right)A\left(\tau\right) \\ -\frac{1}{2}\ddot{A}\left(\tau\right) + i\dot{A}\left(\tau\right) - \mu A\left(\tau\right) = g\chi_0^*\left(\tau\right)\chi_k\left(\tau\right) \end{gathered} \tag{9}$$

where

$$g = eJ\left(8\pi/3\right)^{1/2}\left(N/2V\omega_k^3\right)^{1/2} \tag{10}$$

and

$$\mu = \left(e^2\lambda/\omega_k^2\right)\left(N/V\right) \tag{11}$$

being A the directional averaged vector potential and $\tau = \omega_k t$.

It is easy to show [28,33,35] the differential system (9) admits the following constants of motion

$$\chi_0^*\chi_0 + \chi_k^*\chi_k = 1 \tag{12}$$

$$Q = A^*A + \frac{i}{2}\left(A^*\dot{A} - \dot{A}^*A\right) + \chi_0^*\chi_0 \tag{13}$$

$$H = Q + \frac{1}{2}\dot{A}^*\dot{A} + \mu A^*A + g\left(A^*\chi_k^*\chi_0 + A\chi_0^*\chi_k\right) \tag{14}$$

in which the quantity Q can be considered as the "momentum" of the system and H its Hamiltonian divided by N. In order to study the time – evolution of the system we start from the "perturbative" initial state of QED defined by

$$A\left(0\right) \sim N^{-1/2} \to 0,\ \chi_k\left(0\right) \sim N^{-1/2} \to 0,\ \chi_0\left(0\right) \sim 1 \tag{15}$$

from which the system will "decay" towards the coherent stable state characterized by $A \gg 1$ and $\chi_k \gg 1$. The short-time behavior of the system can be studied [8] by differentiating the third of (9) and substituting it into the second one, so obtaining

$$-\frac{1}{2}\dddot{A}\left(\tau\right) + \ddot{A}\left(\tau\right) + i\mu\dot{A}\left(\tau\right) + gA^2\left(\tau\right) = 0 \tag{16}$$

whose algebraic associated equation is

$$a^3/2 - a^2 - \mu a + g^2 = 0 \tag{17}$$

As know from the general theory, the (17) will have exactly three solutions (real or complex). The "decay" towards the coherent state will occur when the values of μ and g are such to have only one real solution of (17), the other two complex-conjugate ones just describing the exponential increase of $A\left(\tau\right)$ able to overcome its nearly zero initial value and create the coherent tuning field. It can be shown [28-35] that this occurs, for a given μ, when

$$g^2 > g_{crit}^2 \tag{18}$$

where

$$g_{crit}^2 = 8/27 + 2\mu/3 + \left(4/9 + 2\mu^2/3\right)^{3/2} \tag{19}$$

In summary, when condition given by (18) is satisfied, the system will undergo a truly "phase transition" from the incoherent pertubative ground state (PGS) in which the electromagnetic and matter fields perform Zero – Point very weak uncoupled fluctuations only, towards the coherent ground state (CGS) in which a strong electromagnetic field arises from QV and couples with the oscillations of the matter fields tuning all the matter constituents to oscillate in phase with it and among themselves by means of it. But why the should the system do "decide" to run away towards CGS ?

The answer is, as above anticipated, this state is energetically favored so representing the "true" ground state of the electromagnetic field + matter system. This can be rigorously demonstrated by mathematics [28], but a simple physical argument runs as follows.

If we indicate as δE_{ZPF} a spontaneous QED QV fluctuation able to excite some atomic / molecular level of the matter constituents of the given ensemble and with δE_{int} the energy shift induced in them by the interaction with the electromagnetic field of Zero Point (i.e. a Lamb-shift type term), the total energy fluctuation is given by

$$\delta E = \delta E_{ZPF} - \delta E_{\text{int}} \tag{20}$$

where the minus sign before δE_{in} is due to the fact the Lamb-shift term reduce the energy of atomic/molecular constituent, since it introduces in the atomic Hamiltonian the interaction term $e\vec{J}\cdot A(\vec{x},t)$ with

$$\vec{J} = -\sum_{l=1}^{Z} \vec{p}_l / m_e \tag{21}$$

where Z is the atomic number and $\vec{p}_l$ is the momentum operator of the l-th electron. It can be shown [28,30] that, for an ensemble of N atoms/particles interacting with ZPF, we have $\delta E_f \propto N$ while $\delta E_{\text{int}} \propto N\sqrt{N}$ so we can write

$$\delta E = aN - bN\sqrt{N} \tag{22}$$

in which $a > 0$ and $b > 0$ are two constants of proportionality depending on the system properties. From (22) we see that there exist a definite value of $N = N_{crit}$, depending on a and b, such that, when $N \geq N_{crit}$ (namely just the condition for the runaway of the system towards the coherent ground state) we have

$$\delta E < 0 \tag{23}$$

The result given by (23) has a very deep physical meaning since it implies some remarkable consequences [28-35]:

a) the CGS is the "true" ground state of the system because its energy is lower, of the quantity δE (gap), than the energy of "gas-like" PGS in which we only have the independent Zero-Point fluctuations of electromagnetic and matter components while, in the CGS, the matter constituents oscillates in tune with a non fluctuating "strong" electromagnetic field;

b) the "decay" of the system from PGS to CGS can be considered as a truly phase transition, corresponding to the release of a quantity of energy just equal to the gap δE to the environment, so characterizing the electromagnetic field + matter ensemble as an open system;

c) the tuning of the electromagnetic field with the matter field determines a renormalization of frequencies of the matter system so that the common oscillation frequency of electromagnetic field and matter field is given by $\omega_{coh} < \omega_{fluc}$, where $\omega_{fluc} = c^2/\lambda_{CD}$ is the frequency of the QV fluctuating electromagnetic field able to excite the level k and whose wavelength λ_{CD} defines the spatial extension of CD;

d) the coherent electromagnetic field generated inside a CD shows an evanescent tail at its boundary, determining a superposition between the "inner" electromagnetic fields of the neighboring coherence domains. This superposition makes it possible the interaction between different CDs giving rise to the coherence among them also known as "supercoherence" so explaining the physical origin of long-range and stable correlation between a very high number of matter components in living organisms.

As shown by the above discussion, the formation of CD is strictly related to the QV energy density dynamics: the energy needed for the generation of the coherent electromagnetic field is "extracted" from QV (the photons transferred from random quantum fluctuations to tuning electromagnetic field) whose energy density decrease then determining the formation and sustain of the coherent state and the release of the phase transition energy δE to the environment.

III. QED "ELECTRONIC" COHERENCE IN WATER INSIDE BRAIN MICROTUBULES

A. Water critical density inside MT and the transition towards coherent state

As known microtubules are rigid polymers consisting of groups of protofilaments, of length ranging between $1-30\,\mu m$ [26], cylindrically shaped with an outer and inner diameter respectively of about $25\,nm$ and $15\,nm$ (see Fig. 1). They are composed by structural subunits, the tubulin heterodimers (of length about $8\,nm$), in turn containing the $\alpha-$and $\beta-$ tubulin having an high electric dipole moment (about $10^{-26}\,C\cdot m$) [26] and determining the remarkable electric polarity of MTs that make them very sensitive to electromagnetic field. The mechanical properties of MTs have

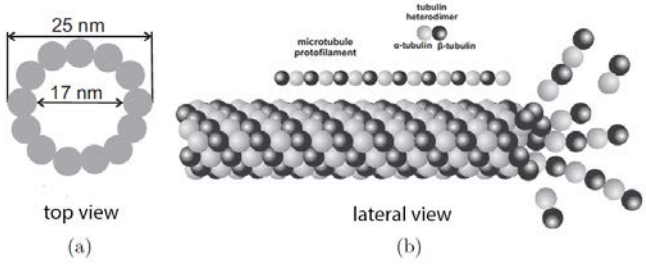


Fig. 1 schematic view of MT structure and its constituents

been studied in details by means of many biophysical techniques such atomic force microscopy, thermal bending and single molecule [36].

Specifically, for the purpose of this paper, we focus on the inner hollow volume of MT that can be assumed to be "filled" with (thermally) isolated water [26], showing in the following that, under the boundary conditions averagely satisfied inside the brain MTs, it undergoes a spontaneous quantum phase transition towards a coherent state in which an electromagnetic field oscillates in tune with the water matter field between two energy levels corresponding to an electronic transition of water molecule.

The coherent dynamics and thermodynamics of liquid water has been analyzed in series of papers [31-35] showing very peculiar and unthinkable features, whose detailed discussion can be found in the cited references. Here we recall some of these that will be specifically used in the present work.

As we have seen above, the "runaway" of a two-levels matter system from PGS to CGS will spontaneously occur when, by (18), $g^2 > g_{crit}^2$, so in order to discover if the considered system will undergo or not the needed superradiant phase

transition from PGS we need to determine the specific value of g_{crit} for our system and the correspondent critical density $\rho_{crit} = m_{H_2O}\left(N/V\right)_{crit}$. To this aim, in the coherent equations (9), the coupling factor g can be written, in the specific case of water, as [33]

$$g = \left(2\pi/3\right)^{1/2}\left(\omega_p/\omega_0\right)^{1/2} f_{01}^{1/2} \tag{24}$$

In which f_{01} is the oscillator strength f_{nk} for the electronic energy transition $0 \leftrightarrow 1$ and ω_p is the plasma frequency given by

$$\omega_p = \left(e/m_e^{1/2}\right)\left(N/V\right)^{1/2} \tag{25}$$

and

$$\mu = -\left(3/2\right)\left(\omega_p/\omega_0\right)^2 \sum_n f_{nk}\,\omega^2 \Big/ \left[\left(E_n - E_k\right)^2 - \omega^2\right] \tag{26}$$

where the oscillator strength for the electronic transition $n \leftrightarrow k$ is given by (6). In [33] the values of g^2, μ and ρ_{crit} related to the first "low-lying" levels of water molecule has been calculated, showing that that smallest value of $\rho_{crit} = 0.310\, g \cdot cm^{-3}$ corresponds to the transition from the ground state to the level at $E = 12.06\, eV$, namely to a $5d$ excited electronic state of water molecule just below the ionization threshold of $12.60\, eV$.

When the water system reaches this density value, the Quantum Vacuum fluctuations with frequency $\omega = \omega_0 = 12.06\, eV$ start to build up coherently with those of the matter field at the same frequency, determining the "runaway" of the system towards the CGS as discussed above. From this point on, the matter + electromagnetic field system behaves as a macroscopic quantum system oscillating with a common frequency ω_{coh} and all the other energetic levels will be totally ignored by the system evolution.

It is now important to note that the value of critical density required for the runaway is compatible with the estimated density of water inside brain MT, $\rho_{water,MT}$. In fact, assuming for the brain an average temperature $T \sim 37\,°C$ and a MT cavity volume [26] $V_{MT} \sim 5 \cdot 10^{-22}\, m^3$ for a moderately long ($l \sim 10^{-6}\, m$) MT, we have

$$\rho_{water,MT} \sim 0{,}993\, g \cdot cm^{-3} > \rho_{crit} \tag{27}$$

showing the existence of the conditions, inside MT inner volume, required for the superradiant phase transition of water towards the coherent state.

The key point to stress now is that the energy level $\omega_0 = 12.06\, eV$ correspond to a CD whose "size" is, by (2) of order of

$$L_{CD} \sim 0.1\,\mu m \tag{28}$$

that is about $1/10$ of the average length of a moderately sized MT and, in particular, of the order of magnitude of the MT dimers ($\sim 8\, nm$).

The rough estimate given by (28) is very important since it shows that superradiant photons, generated in coherent electromagnetic field oscillating in phase with water molecules inside MT coupled to the electronic transition from the ground state to the energetic level at 12.06 eV, are characterized by a wavelength much shorter than the typical length of MT so allowing the formation of node and antinodes within the inner MT cavities.

This very important result completely overcome the actual trouble raised by Georgiev with respect the infrared superradiant photons considered in the theoretical models of coherence in MT presented so far.

As recalled in the above discussion, the coherent dynamics inside CD determines a rescaling of the frequency ω_{coh} of the common oscillation of electromagnetic field and matter to a lower value with respect that of ω_0 characterizing the perturbative state in which they are out of phase. It has been shown [28,32,33,35] the "new" value of frequency to be

$$\omega_{coh} = \left|1 - \dot{\phi}\right|\omega_0 \tag{29}$$

where ϕ is the phase factor ruling the behavior of the vector potential $A(\tau) = A_0 \exp\left[i\phi(\tau)\right]$. It is interesting to note that, in the case of water [28,32-35], $\omega_{coh} \sim 10^{-2}\omega_0$, determining an energy gap per molecule $\delta E/N \sim -0.26\, eV$.

B. About the photon mass value in the water coherence domains inside MT and the evanescent field of superluminal photons

One of the most important consequences of coherent dynamics, deriving from the frequency rescaling of (29), is that the superradiant photon "mass" acquires an imaginary value inside the coherent electromagnetic field. This can be easily seen by using (2) and the Einstein equation for a the photon, obtaining

$$\begin{aligned} & m^2c^4 = \\ & = \hbar\left(\omega_{coh}^2 - 4\pi^2c^2/\lambda_{CD}^2\right) < \hbar\left(\omega_0^2 - 4\pi^2c^2/\lambda^2\right) = 0 \end{aligned} \tag{30}$$

so implying $m = i \cdot m_0$, where m is the photon mass inside the CD, namely just the condition, given by (5), associated to the existence of superluminal photons ! Now we have obtained our second important result: inside the CDs originated in brain's MTs by the coherent dynamics of water, the superradiant photons, populating the coherent electromagnetic field tuned with matter field, can be considered as moving at a superluminal velocity inside the CD itself. Furthermore, the condition (30) also states that the above electromagnetic field is "trapped" inside the CD, thus preventing its dissolution by radiating the coherent field

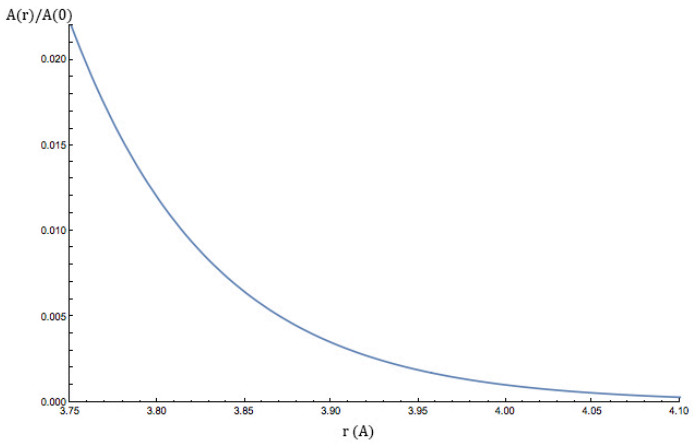


Fig. 2 tail of coherent electromagnetic field of a water CD

towards outside.

Now, it is then interesting to better investigate the spatial behavior of the coherent solution for the electromagnetic field given by (9). As derived in [28,33] the spatial dependence of the electromagnetic amplitude $A(\vec{x},t)$ can be given, considering the spherical symmetry of the problem, by the following equations

$$\begin{cases} A(\vec{x},t) = A(0)\left(\sin\omega_0 r / \omega_0 r\right)\cdot\exp\left(-i\omega_{coh}t\right),\ r < r_{coh} \\ \left(-\omega_{coh}^2 - \nabla^2\right)A(\vec{x}) = 0,\ r > r_{coh} \end{cases} \tag{31}$$

where r_{coh} is the "radius" of the CD and the second equation means that, outside the CD, $A(\vec{x},t)$ satisfies a free field equation. Now using the fact that $\nabla^2 A \sim \omega_0^2 A(r)$ we can write

$$\left(d^2/dr^2\right)(r\cdot A) - \left(\omega_0^2 - \omega_{coh}^2\right)(r\cdot A) = 0 \tag{32}$$

that, as can be easily seen, is characterized by an exponential decaying solution

$$A(r) \sim \exp\left(-r\sqrt{\omega_0^2 - \omega_{coh}^2}\right) \tag{33}$$

showing the presence of an "evanescent" electromagnetic field at the borders of the CD. In particular, by imposing the matching, at the CD boundary $r = r_{coh}$, between the exponential solution given by (33) and the first of (31) and recalling that $\omega_0 \gg \omega_{coh}$, we obtain

$$r_{coh} \simeq 3\pi/4\omega_0 \tag{34}$$

that represents a better estimate of the dimension of CD. For the case of water with $\omega_0 = 12.06\,eV$ we have

$$r_{coh} \sim 3.75\cdot 10^{-8}\,m \tag{35}$$

that further confirm our previous result of (28) showing, in particular, that the cavity inside MT can be though as "filled" with water CDs associated to the coherent dynamics related to the electronic transition from the ground state to the level at $12.06\,eV$. The superradiant "evanescent" field is then given by

$$\begin{aligned} A(r) &\simeq \left(A(0)/\sqrt{2}\right)\times \\ &\times\left(\exp\left[-\sqrt{\omega_0^2 - \omega_{coh}^2}\,(r - r_{coh})\right]/\omega_0 r\right) \end{aligned} \tag{36}$$

whose profile is shown in Fig. 2. Another very meaningful consequence of the above result is that the coherent electromagnetic field resulting from the tuned interaction between matter and electromagnetic field inside the CD has a "tail" extending outside it, under the form of evanescent field, whose spatial extension makes it able to overlap the electromagnetic field associated to the neighborhood CDs.

According to this mechanism, contiguous CDs can interact each other realizing the long-range correlation need for the implementation of biological functions. In particular, this tail allows the evanescent electromagnetic field associated to the water CDs inside MT to "cross" the MT wall and interact with the biological structures placed on it and in its neighborhood.

The existence of this "evanescent" electromagnetic field, emerging from the water CDs, then theoretically suggests, on a robust QFT basis, a possible physical mechanism able to explain the tunneling of superluminal photons, trapped inside water CD, towards the "outside" environment.

C. Thermodynamics of water inside brain MT and the environmental decoherence problem

In the coherent state so far analyzed, the tuned oscillation between matter and electromagnetic field forbids any thermal fluctuation and then it is virtually associated with a thermodynamic absolute temperature $T = 0$. In this condition, no energy inflow from the environment is then possible. This is prevented by the energy gap characterizing the coherent state after the release of the energy δE .

Nevertheless, if the temperature of the environment increases to a value $T > 0$ (as, for example, occurs for CD placed in a thermal bath at $T \neq 0$), the collisions between the fluctuating environment molecules (thermally excited) and the components of a CD, could transfer to it the energy gap per atom/molecule δE , able to put some of them out of tune with the electromagnetic field.

This environmental decoherence determines the "expulsion" of some matter components from the CD and the formation of an incoherent fraction of matter system at the boundaries of CD.

So, in order to ensure the formation and the persistence of CDs in the water inside brain MT, the fraction of coherent water has to be sufficiently high.

In order to verify the occurrence of this condition in brain MT, we'll consider the expression of coherent water fraction F_{coh} as a function of absolute temperature T obtained in [28,33], namely

$$F_{coh}(T) = 1 - \left(64/9\right)\int_0^{3/4} x^2 F(x,T)\,dx \tag{37}$$

where

$$F(x,T) = Z\cdot\exp\left[-\delta E(x)/T\right] \tag{38}$$

$\delta E\left(x\right)$ is the energy gap of the coherent state as a function of the distance x from the CD centre and Z is the partition function [28,33] given, in this case, by

$$Z = \left(N/V\right)\left(m \cdot T/2\pi\right)^{1/2}\left(k^2/2\pi\right)\cdot \exp\left(-\delta_0/T\right) \quad (39)$$

where m is the mass of a water molecule and, in the case of water, $k \sim 5 \cdot 10^{-10}\, m$ and $\delta_0 \sim 400\, cm^{-1}$.

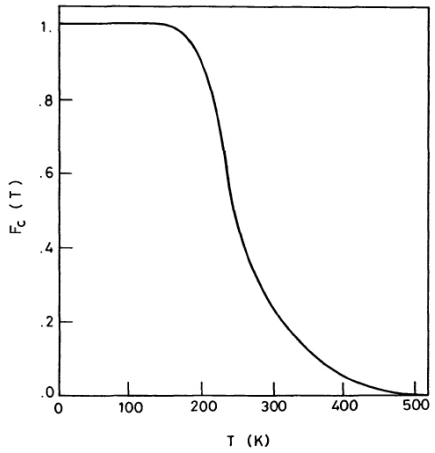


Fig. 3 behavior of coherent fraction in water [28,33]

The behavior of $F_{coh}\left(T\right)$ for bulk water is represented in the Fig. 3, from which we deduce that, for a temperature of about $T \sim 310\, K$ corresponding to the average temperature inside brain MT, we should have a coherent fraction $F_{coh}\left(T_{brain}\right) \sim 0.2$ that is approximately less than a half of the corresponding coherent fraction at room temperature. Nevertheless for the water enclosed within a cavity, as happens in MTs, it has been shown [37] that cavity wall is able to decrease the impact of thermal fluctuations so making the interfacial water substantially thermally isolated and then much more coherent than bulk water.

Since practically all the water contained in a living organisms is always very near to a "wall" [37] (typically less than a fraction of micron from a surface like a membrane or a molecular backbone) we can consider this water as interfacial water and then, for the above considerations, we can assume $F_c \to 1$.

A further confirmation of this assumption results from the experimental evidence that water inside cells resides in a sort of "glassy" state [38] whose coherent general properties has been already investigated [39] showing that, for this water, the coherent fraction is the same of that occurring for $T < 220\, K$ namely, from the Fig. 3, $F_c \sim 1$ able to guarantee the existence and the permanence of an almost fully-coherent state of water inside brain MTs.

In this way we have shown that also the question of environmental decoherence, properly raised by Mavromatos et al [26], can be issued within the framework of QED coherence in water, when we consider the coupling between electromagnetic and matter field occurring at the level of electronic energy transitions.

IV. Possibility of Hypercomputation in Brain MT by Means of Superluminal Photons

Feynmann defined the required energy per step for the computation process given by [40]

$$energy\ per\ step = 2k_B T\left(f-b\right)/\left(f+b\right) \quad (40)$$

where k_B is Boltzmann's constant, T is a temperature, f is a forward rate of computation and b is backward rate.

Supposing that there in no energy supply and parameters f and b are fixed during the computation, we can consider the infinite computational steps given by

$$E_1 = kE_0, E_2 = kE_1, ..., E_n = kE_{n+1}, ... \quad (41)$$

where we let the initial energy of computation be $E_0 = k_B T$, $k = 2\left(f-b\right)/\left(f+b\right)$ and E_n is the energy for the n-th step computation.

From the above we have $E_n = k^n E_0$, and then the energy loss for each computational step becomes

$$\begin{aligned} \Delta E_1 &= E_0 - E_1 = \left(1-k\right)E_0 \\ \Delta E_2 &= E_1 - E_2 = \left(1-k\right)kE_0 \\ &\vdots \\ \Delta E_n &= E_{n-1} - E_n = \left(1-k\right)k^{n-1}E_0 \end{aligned} \quad (42)$$

According to the paper by S. Lloyd [41], it is required for the quantum system with average energy ΔE to take time at least Δt to evolve to an orthogonal state given by

$$\Delta t = \pi\hbar/2\Delta E \quad (43)$$

from which, the total energy for the infinite steps yields E_0 if setting $E = \Delta E_i$ into (43), then the total time for the computation with infinite steps becomes

$$T = \sum_{n=1}^{\infty}\Delta t_n = \left(\pi\hbar/2E_0\right)\sum_{n=1}^{\infty} 1/\left(1-k\right)k^{n-1} \quad (44)$$

As the infinite sum of (44) diverges to infinity when satisfying $0 < k < 1$, the Feynman model of computation requires infinite time to complete the calculation.

Hence an accelerated Turing machine cannot be realized for computers utilizing ordinary particles due to the uncertainty principle. Recami claimed in his paper [20] that tunneling photons which travel in evanescent mode can move with superluminal group speed inside the barrier.

From the relativistic equation of energy and momentum of the moving particle, given by

$$E = m_0 c^2 \Big/ \sqrt{1-\left(v/c\right)^2} \quad (45)$$

and

$$p = m_0 v \Big/ \sqrt{1-\left(v/c\right)^2} \quad (46)$$

the relation between energy and momentum can be shown as $p/v = E/c^2$.

From which, we have [42]

$$\left(v\Delta p - p\Delta v\right)/v^2 = \Delta E/c^2 \tag{47}$$

Supposing that the approximation $\Delta v/v^2 \approx 0$ holds, the (47) can be simplified as

$$\Delta p \approx \left(v/c^2\right)\Delta E \tag{48}$$

This relation is also valid for the superluminal particle (which has an imaginary mass $i \cdot m_0$), the energy and the momentum of which are given by following equations, respectively

$$E = m_0 c^2 \Big/ \sqrt{\left(v/c\right)^2 - 1} \tag{49}$$

$$p = m_0 v \Big/ \sqrt{\left(v/c\right)^2 - 1} \tag{50}$$

According to the paper by M.Park and Y.Park [43], the uncertainty relation for the superluminal particle can be given by

$$\Delta p \cdot \Delta t \approx \hbar/\left(v - v'\right) \tag{51}$$

where v and v' are the velocities of a superluminal particle after and before the measurement. By substituting (48) into (51), we obtain the uncertainty relation for superluminal particles given by

$$\Delta E \cdot \Delta t \approx \hbar/\beta(\beta - 1) \tag{52}$$

when we let $v' = c$ and $\beta = v / c$.

Instead of subluminal particles including photons, the time required for the quantum system utilizing superluminal particles becomes [44]

$$\begin{aligned} T &= \sum_{n=1}^{\infty} \Delta t_i = \\ &= \left(\pi\hbar/2E_0\right)\sum_{n=1}^{\infty} 1/\beta_n\left(\beta_n - 1\right)\left(1 - k\right)k^{n-1} \end{aligned} \tag{53}$$

from the uncertainty principle for superluminal particles given by (52), where β can be given by

$$\beta_n = \sqrt{1 + m_0^2 c^4 / E_n^2} = \sqrt{1 + m_0^2 c^4 / k^{2n} E_0^2} \tag{54}$$

which is derived from (50).

From (53) and (54), it is seen that the computation time can be accelerated as shown in Fig. 4.

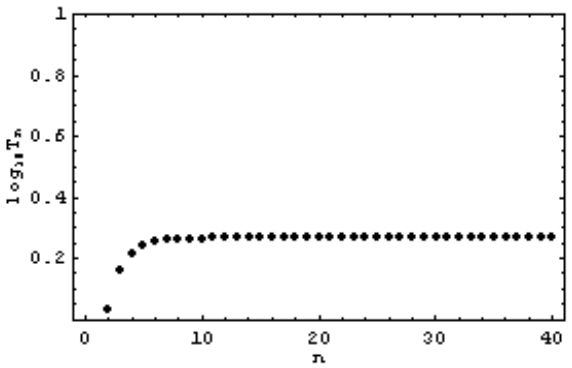


Fig. 4 time required to conduct computation at each step by using superluminal particles (for the case, $k = 1/2$ and $\gamma = 1.0$).

By the numerical calculation, it can be shown that the infinite sum of (53) converges to a certain value satisfying $0 < k < 1$, as shown in Fig. 5.

In this figure, the horizontal line is for the parameter $\gamma = m_0 c^2 / E_0$ and the vertical line is for the time to complete infinite step calculations. This result means that infinite steps of computation can be conducted within a finite length of time by using superluminal particles.

From these calculation results, it can be seen that a hypercomputing machine can be realized by utilizing superluminal particles, instead of subluminal particles, for the Feynman's computational model.

Thus, contrary to the conclusion obtained relatively to the Feynman's model of computation when using ordinary particles, it can be seen that superluminal particles permits the realization of an accelerated Turing machine.

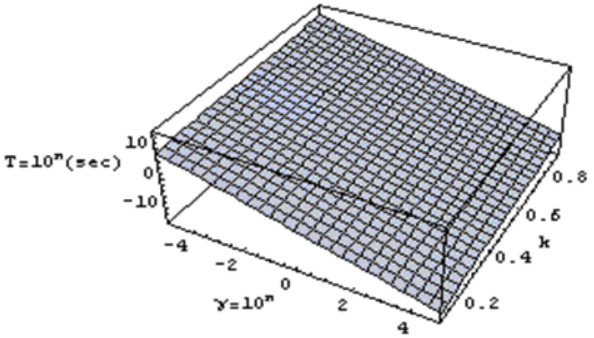


Fig. 5 computational required time for the superluminal particles.

It is known that an accelerate Turing machines allow us to be computed some functions which are not Turing-computable, such as the halting problem [45], described as "given a description of an arbitrary computer program, decide whether the program finishes running or continues to run forever".

Thus superluminal photons inside the microtubule permit us to conduct hypercomputation which cannot be realized by the conventional silicon processors.

V. Conclusion

In this paper we have shown, basing on the theoretical framework of QED coherence in condensed matter, that water contained inside the hollow volume of brain microtubules is

able to exhibits a spontaneous superradiant quantum phase transition toward a energetically favored state, in which the electronic clouds of water molecules coherently oscillate in tune with a self-trapped electromagnetic field within defined space regions (coherent domains).

Differently from the models of quantum optical coherence in cytoskeletal MT proposed so far (that considers the energetic transitions of water molecules associated to rotational energy levels of the order of few meV), the picture here discussed assume the coherent system (water + electromagnetic field) to oscillate in phase with the electronic transition of water from the ground state to the level at energy $E = 12.06\,eV$, implying the supperradiant photons, generated inside the coherence domains, to have a wavelength much smaller than the length of a moderately sized MT in brain.

Furthermore, the coherent electromagnetic field arising from quantum vacuum oscillating in tune with water molecules is characterized by a negative squared mass of the superradiant photons (since they are trapped inside the coherence domain) and by an "evanescent" tail extending outside the coherence domain itself. In particular, these two latter features allows us to interpret these photons as superluminal evanescent (tunneling) photons that, as it has been shown in this paper, can be used by living system to implement high performance quantum computing inside brain using microtubules substrate as storage material.

The proposed model also overcomes some of the most important issues (as, in first place, the too long wavelength of superradiant photon with respect MT size and the too short environmental decoherence time) properly raised by some authors about the actual possibility to consider the superrandiant photons, generated inside MT from water coherent phase transition, in order to explain the occurrence of "ordinary" functions performed by brain as well as its eventual quantum hypercomputing capabilities.

Obviously, further researches are needed in order to fully understand the above mechanisms and, to this aim, different further aspects of coherent interaction between microtubule structures and water they contain and their implications on brain capabilities, functions and processes as, for example, consciousness, will be analyzed in forthcoming publications.

REFERENCES

[1] W. Meissner, R. Ochsenfeld, "Ein neur effect bei eintritt der supraleitfahigkeit", Naturwissenschaften, vol. 21, n. 44, pp. 787-788, 1933.

[2] F. London, H. London, "The electromagnetic equations of the superconductor", *Proceedings of the Royal Society of A: Mathematical, Physical and Engineeering Sciences*, vol. 149, 1935.

[3] V. L. Ginzburg, L. Landau, *Zh. Eksp. Teor. Fiz.*, 20(1064), 1950.

[4] J. Barden, L. N. Cooper, J. R. Schrieffer, "Theory of superconductivity", *Phys. Rev*. vol. 106, pp. 162-164, 1957.

[5] I. N. Marshall, "Consciousness and Bose-Einstein condensates", *New Ideas Psychol*., vol. 7, pp. 73-83, 1989.

[6] F. Crick, C. Koch, "Towards a neurobiological theory of consciousness", *Semin. Neurosci*, vol. 2, pp. 263-275, 1990.

[7] W. Singer, "Synchronization of cortical activity and its putative role in information processing and learning", *Annu. Rev. Physiol*., vol. 55, pp. 349-374, 1993.

[8] R. Penrose, "*The emperor's new mind*", Oxford: Oxford University Press, 1989.

[9] R. Penrose, "*Shadows of the mind*", Oxford: Oxford University Press, 1994.

[10] K. H. Pribram, "*Brain and Perception*", New Jersey: Lawrence Erlbaum, 1991.

[11] M. Jibu et al., "Quantum optical coherence in cytoskeletal microtubules: implication for brain function", *BioSystems*, vol. 32, pp. 195-209, 1994.

[12] E. Schrodinger, "What is life ?", Cambridge: Cambridge University Press, 1944.

[13] H. Frohlich, "Long-range coherence and energy storage in biological systems", *Int. J. Quantum Chem.*, vol. 2, n. 5, pp. 641-649, 1968.

[14] F. A. Popp, "*Recent advances in biophoton research and its application*", Singapore, London, New York: World Scientific, 1992.

[15] H. Umezawa, "*Advance field theory: micro, macro and thermal physics*", New York: American Institute of Physics, 1993.

[16] A. S. Davydov, "Solitons in molecular systems", *Phys. Scripta*, vol. 20, pp. 387-394, 1979.

[17] S. R. Hameroff, R. Penrose, "Conscious events as orchestrated space-time selections", *J. Consciousness Studies*, vol. 2, 1995.

[18] S. R. Hameroff, R. C. Watt, "Information processing in microtubules", *J. Theor. Biol*., vol. 98, pp. 549-561, 1982.

[19] E. Del Giudice, G. Preparata, G. Vitiello, "Water as a free electric dipole laser", *Phys. Rev. Lett.*, vol. 61, pp. 1085-1088, 1988.

[20] E. Recami, "A bird's-eye view of the experimental status-of-the-art for superluminal motions", *Found. Phys*., vol. 31, pp. 1119-1135, 2001.

[21] E. Recami, "Superluminal tunneling through successive barriers: does QM predict infinite group velocities ?", *J. Modern Opt*., vol. 51, pp. 913-923, 2004.

[22] T. Musha, "Possibility of high performance quantum computation by superluminal evanescent photons in living systems", *BioSystems*, vol. 96, pp. 242-245, 2009.

[23] L. M. Caligiuri, A. Sorli, "Space and time separation, superluminal motion and Big Bang cosmology", *Journal of Cosmology*, vol. 18, pp. 212-222, July/ August 2014.

[24] R. W. Ziolowski, "Superluminal transmission of information through an electromagnetic material", *Phys. Rev. E*, vol. 63, n. 4, 2001.

[25] D. D. Georgiev, "Quantum computation in the neural microtubules: quantum gates, ordered water and superradiance", arXiv:quant-ph/0211080.

[26] N. E. Mavromatos, "Quantum coherence in (brain) microtubules and efficient energy and information transport", *Journal of Physics: Conference Series*, vol. 329, pp. 1-31, 2011.

[27] L. M. Caligiuri, "Zero-Point field, QED coherence, living systems and biophotons emission", *Open Journal of Biophysics*, vol. 5, n. 1, January 2015, to be published.

[28] G. Preparata, "*QED coherence in matter*", Singapore, London, New York: World Scientific, 1995.

[29] E. Del Giudice, S. Doglia et al., "*Structures, correlations and electromagnetic interactions in living matter*", Berlin: Springer-Verlag, 1988.

[30] E. Del Giudice, R. Mele et al., "Hamiltonian and superradiant pahse transition", *Mod. Phys. Lett. B*, vol. 7, pp. 1851-1855, 1993.

[31] E. Del Giudice, G. Preparata, "Coherent dynamics in water as possible explanation of membrane formation", *J. of Biol. Phys*., vol. 20, pp. 105-116, 1994.

[32] E. Del Giudice, P. Preparata, "*A new QED picture of water: understanding a few fascinating phenomenon. Macroscopic quantum coherence*", E. Sassaroli, Y. Strivastava et al. eds., Singapore, London, New York: World Scientific, 1988, pp. 108-129.

[33] R. Arani, I. Bono, E. Del Giudice, G. Preparata, "QED coherence and the thermodynamics of water", *Int. J. of Mod. Phys*., vol. 9, n. 15, pp. 1813-1841, 1995.

[34] V. L. Voeikov, E. Del Giudice, "Water respiration: the basis of the living state", *Water*, vol. 1, pp. 52-75, 2009.

[35] I. Bono, E. Del Giudice, L. Gamberale et al., "Emergence of the coherent structure of liquid water", *Water*, vol. 4, pp. 510-523, 2012.

[36] M. Cifra, J. Pokorny, D. Havelka et al, "Electric field generated by axial longitudinal vibration modes of microtubules", *BioSystems*, vol. 100, pp. 122-131, 2010.

[37] E. Del Giudice, P. Stefanini, A. Tedeschi et al., "The interplay of biomolecules and water at the origin of the active bahaviour of living organisms", *Journal of Physics: Conference Series*, vol. 329, 2011.

[38] X. Trepat, L. Deng, S. An et al, "Universal physical response to stretch in the living cell", *Nature*, vol. 447, pp. 592-595, 2007.

[39] M. Buzzacchi, E. Del Giudice, G. Preparata, "Coherence of the glassy state", *Int. J. Mod. Phys. B*, vol. 16, 2002.
[40] R.P.Feynman, "*Feynman Lectures on Computation*", London: Penguin Books, 1996.
[41] S.Llyod, "Ultimate physical limit to computation", *Nature*, vol.406, pp. 1047-1054, 2000.
[42] T.Musha, "Possibility of Hypercomputation by Using Superluminal Elementary particles", *Advances in Computer Science and Engineering*, vol 8, n. 1, pp. 57-67, 2012.
[43] M. Park, Y. Park, "On the foundation of the relativistic dynamics with the tachyon", *Nuovo Cimento*, vol.111B, n.11, pp. 1333-1368, 1996.
[44] T.Musha, "Possibility of Hypercomputation from the Standpoint of Superluminal Parrticles", *Theory and Applications of Mathematics & Computer Science*, vol. 3, n. 2, pp. 120-128, 2013.
[45] T.D.Kieu, "Hypercomputation with quantum adiabatic processes", *Theoretical Computation Science*, vol. 317, pp. 93-104, 2004.

On the kinetics of biogenic amines formation under different levels of selected factors

M. Tláskal, F. Buňka, J. Michálek, L. Buňková, and P. Pleva

***Abstract*—**Some bacterial strains of enterococci are commonly used in food industry and therefore their ability of biogenic amine formation should be investigated. This enables to indicate decarboxylase-positive strains. Within the process of decarboxylation, these strains produce high amount of biogenic amine, which is a toxicologically important compound. Biogenic amines are present in certain foodstuffs (cheese, meat, wine ...) and at high concentrations they are considered as risk factors for human health. The aim of this contribution was to explore production of eight chosen biogenic amines by *Enterococcus faecium* (DPE 002) from rabbit meat (*Oryctolagus cuniculus f. domesticus*) and to evaluate the effect of selected factors on the production.

To fit the data subsets involving different conditions of the experiment, appropriate regression models were used. Some of the growth curves such as Gompertz, logistic, and Richards are found to be very useful in many areas. The most suitable models for our data appeared to be Gompertz and logistic. Their three regression parameters, which are of biological interest, are an asymptotic value of concentration, a maximal growth rate and a lag time. Model parameters were estimated and the effect of different factor levels on the parameter values is studied.

***Keywords*—**Biogenic amines, Gompertz curve, Growth model, Logistic curve.

I. Introduction

BIOGENIC amines (BAs) are basic compounds (especially histamine, phenylethylamine, tyramine, tryptamine, putrescine, cadaverine, spermidine and spermine) formed in foodstuff mainly by microbial decarboxylation of relevant free amino acids (especially histidine, phenylalanine, tyrosine, tryptophan, lysine, ornithine and arginine). Many strains *Salmonella*, *Shigella*, *Escherichia*, *Serratia*, *Yersinia*, *Morganella*, *Pseudomonas* and lactic acid bacteria (e. g. *Lactobacillus* and/or *Enterococcus*) were identified as producers of BAs. The presence of biogenic amines is usually connected with food poisoning and can thereby threaten health of its consumers. Ordinary amounts of BA (practically <100 mg/l or mg/kg) in food are metabolized in intestinal tract of healthy individuals where they are detoxicated by proper enzymes (especially monoamine oxidase and diamine oxidase). Especially in sensitive consumers, excessive intake of histamine and tyramine can result in hyper- or hypotension, migraine, headache, vomiting, and respiration problems. Putrescine and cadaverine can potentiate the impact of tyramine and histamine occurrence, because they inhibit their detoxication enzymes. Polyamines (especially spermine and spermidine) can be converted to carcinogenic nitrosamines, which represent an issue to be studied [4], [8], [9].

Since some strains of enterococci are even used as starter cultures and/or probiotics, their decarboxylase activity should be studied to indicate decarboxylase-positive strains. The aim of the study was to explore production of eight biogenic amines (tryptamine, phenylethylamine, histamine, cadaverine, tyramine, putrescine, spermine and spermidine) by the selected *Enterococcus faecium* (DPE 002) from rabbit meat (*Oryctolagus cuniculus f. domesticus*) [7]. Different levels of factors influencing decarboxylase activity (pH, oxygen availability, concentration of NaCl and temperature) were set up in the experiment.

Mathematical models describing laws of growth and development phenomena are needed in many fields, like e.g. food microbiology (see [1], [3], [11]), crop science, forestry, and animal science. We have considered Gompertz, logistic and Richards model (see [11]).

II. Methods and Materials

Effects of additions of NaCl, values of pH and aerobic/anaerobic conditions, that could influence production of biogenic amines was tested using *Enterococcus faecium* (isolated from rabbit meat (*Oryctolagus cuniculus f. domesticus*)). The tested *Enterococcus faecium* (DPE 002) strain was cultivated in MRS broth (Oxoid, Basingstoke, UK) enriched with the precursors of the monitored BAs (amino acids: arginine, histidine, lysine, ornithine and tyrosine, each with the concentration of 2 $g \cdot L^{-1}$; Sigma-Aldrich, St. Louis, USA). The cultivation medium of the volume 5 ml was inoculated with 25 µl overnight culture of the strain

This work was supported by the specific research project "Modelling of risk phenomena" (SV14-FEM-K101-01-MICH).

M. Tláskal and J. Michálek are with the Department of Econometrics, Faculty of Military Leadership, University of Defence, Kounicova 65, 612 00 Brno, Czech Republic (e-mail: martin.tlaskal@unob.cz; jaroslav.michalek@unob.cz).

F. Buňka is with the Department of Food Technology, Faculty of Technology, Tomas Bata University in Zlin, nam. T. G. Masaryka 5555, 760 01 Zlin, Czech Republic (e-mail: bunka@ft.utb.cz).

L. Buňková and P. Pleva are with the Department of Environmental Engineering Protection, Faculty of Technology, Tomas Bata University in Zlin, nam. T. G. Masaryka 5555, 760 01 Zlin, Czech Republic (e-mail: bunkova@ft.utb.cz; ppleva@ft.utb.cz).

(~10^6 CFU/ml).

Experimental setup to the following scheme: (i) effect of NaCl additions at concentrations 0; 10; 20; 30 and 60 g/L; (ii) effect of pH-value: 5.0; 6.0 and 7.0; (iii) cultivation temperature: 6; 12 and 30 °C; and (iv) aerobic and anaerobic environment. The development of biogenic amines was observed during cultivation: 0; 2; 4; 6; 8; 10; 12; 24; 30; 34 and 48 hours (30 °C), 0; 24; 48; 72; 96; 144; 168; 216; 240; 312 and 360 hours (12 and 6 °C).

The production of eight biogenic amines (tryptamine, TRY; histamine, HIS; tyramine, TYR; phenylethylamine, PHE; putrescine, PUT; cadaverine, CAD; spermidine, SPD; spermine, SPN) was monitored by an high performance liquid chromatography system equipped with a binary pump; an autosampler (LabAlliance, State College, USA); a column thermostat; a UV/VIS DAD detector (λ = 254 nm); and a degasser (1260 Infinity, Agilent Technologies, Santa Clara, USA).

After cultivation of the tested bacteria, the broth was centrifuged (4000 x *g*; 22±1 °C; 20 minutes) and the acquired supernatant was diluted (1:1; v/v) with perchloric acid (c = 0.6 mol.l^{-1}). The mixture was filtered (porosity 0.22 μm) and the acquired filtrate was subjected to derivatisation with dansylchloride according to [2]; 1,7-heptanediamine was used as an internal standard. The derivatised samples were filtered (porosity 0.22 μm) and applied on a column (Agilent Zorbax Eclipse C18, 50 × 3.0 mm, 1.8 μm; Agilent Technologies; Agilent, Paolo Alto, CA, USA). The conditions for separation of the monitored BA are described by [6]. Each of the three cultivated broths was derivatised once and applied on the column.

To sum up, totally 33 measurements were made (11 time points, 3 replicates) for every combination of individual factor levels and the total number of combinations was 90.

III. Statistical analysis

The three nonlinear regression models that we have considered as suitable for description of temporal progress of biogenic amine concentration were Gompertz, logistic and Richards model. Equations of the sigmoidal growth curves, which correspond to these models are as follows:

Gompertz model:

$$y(t)= A \exp\left\{- \exp\left[\frac{\mu \cdot e}{A}(\lambda - t)+1\right]\right\}$$

logistic model:

$$y(t)= A \left\{1+ \exp\left[\frac{4\mu}{A}(\lambda - t)+2\right]\right\}^{-1}$$

Richards model:

$$y(t)= A \left\{1+ \nu \exp\left[1+\nu+\frac{\mu}{A}(1+\nu)^{1+1/\nu}(\lambda - t)\right]\right\}^{-1/\nu}$$

Parameterization of the curves according to [11], where their parameters have a clear biological meaning, was used. The parameter A is the asymptotic concentration (for *t* approaching ∞, in mg·l^{-1}), μ is the maximal production rate (in mg·l^{-1}·h^{-1}) and λ is the lag time (in hours, defined as the *t*-axis intercept of the tangent through the inflection point).

All the three curves under consideration are S-shaped and have some similar properties [10]. On the other hand, a substantial difference between them is the ordinate of the point of inflection. In the case of Gompertz curve it is A/e, in the case of logistic curve it is A/2, and in the case of Richards curve, the additional parameter ν affects the ordinate of the inflection point [11].

Every data set representing different conditions of the experiment was processed by tools of nonlinear regression analysis. As appropriate nonlinear models we have considered Gompertz, logistic and Richards model. By means of Akaike information criterion it was decided which model fits the data best. Where possible, the datasets were tried to be fitted by the three above mentioned models. Nevertheless, for some datasets none of the models was considered as suitable and therefore model-free spline fit was used for description of the data. In these cases, standard deviations of parameters were not calculated and the combination of experimental conditions was not encompassed in the subsequent Bonferroni procedure. Regression analysis was performed with the use of the R 3.0.1 software, the package *grofit* was exploited for fitting data sets by growth curves. To obtain initial values of regression parameters (A, μ, and λ), the given time series was fitted by local weighted regression method (implemented in the function *lowess*). For more information about *grofit* package and corresponding R-functions that are used in connection with the problem of growth curve fitting see [5]. The package is available from http://cran.r-project.org/package=grofit. A cubic interpolation spline and its values of the parameters were used as initial parameter estimates for the subsequent iterative procedure. Final estimates of parameters were obtained by applying Gauss-Newton method.

The parameter estimates and their standard deviations were calculated and under different experimental conditions, differences in their values became obvious. To test whether the observed differences are statistically significant, method of multiple comparisons was applied. We have assumed an asymptotic normality of estimated parameters. Multiple comparisons of parameters were performed by the Bonferroni method (α = 0.05). This procedure was carried out for all datasets excepting the sets that were modelled by spline.

IV. Results of analysis

None of the datasets was fitted by Richards model. It seems that the three parameter Gompertz and logistic curves are models, which are sufficient for the description of kinetics of biogenic amines formation. Production of high concentrations was detected in the cases of phenylethylamine and tyramine.

Detailed results follow.

A. Results for phenylethylamine formation

Growing pattern was observed for temperatures of 12 and 30 °C. In these cases, coefficient of determination R^2 realised in the range from 0.721 to 0.999. For temperature 6 °C the maximal mean concentration (obtained from spline fit) is less than 10 $mg \cdot l^{-1}$ and decreases with increasing pH value. The maximum asymptotic concentration determined by the parameter A was observed for temperature 30 °C, pH value 6, concentration of NaCl 2 % and anaerobic environment.

The effect of experimental conditions on value of parameter A

The impact of pH of the environment is such that substantially lower asymptotic concentrations were provided in the case of pH equal to 7. As a general rule, in the case of the highest NaCl concentration 6 %, values of parameters A were higher in the environment of pH equal to 5 than equal to 6. Nevertheless, for lower values of NaCl concentration (0, 1, 2, and 3 %) we get an opposite inequality. These conclusions are valid for both aerobic environment and anaerobic environment. The asymptotic concentrations grew higher under anaerobic conditions than in aerobic environment in the case of temperature 30 °C and pH value equal to 6 and 7. For the pH value of 5, no such phenomenon is evident. The influence of NaCl concentration is not so clear, and thus we cannot make any general conclusions. At a significance level of 5 % all the factors proved to be statistically significant. More precisely, the hypothesis of A parameters equality was rejected in 100, 95, 79, and 84 % of all cases, respectively for the factor temperature, pH, NaCl concentration, and aerobic/anaerobic environment, respectively.

The effect of experimental conditions on value of parameter μ

Only the impact of pH of the environment is apparent: the maximum growth rate was highest in the case of pH equal to 6 and lowest in the case of pH equal to 7. This holds for both aerobic environment and anaerobic environment, as well as for arbitrary chosen level of NaCl concentration. At a significance level of 5 % all the factors proved to be statistically significant. The hypothesis of μ parameters equality was rejected in 63, 90, 83, and 89 % of all cases, respectively for the factor temperature, pH, NaCl concentration, and aerobic/anaerobic environment factor, respectively.

The effect of experimental conditions on value of parameter λ

The parameter λ - lag time takes greater value for the level of temperature 12 °C than for 30 °C. This holds true for all levels of the other factors. The influence of all the factors was evaluated as statistically significant again (at a significance level of 5 %). The hypothesis of λ parameters equality was rejected in 100, 70, 76, and 53 % of all cases, respectively for the factor temperature, pH, NaCl concentration, and aerobic/anaerobic environment factor, respectively.

The impact of pH of the environment can be seen in Fig. 1.

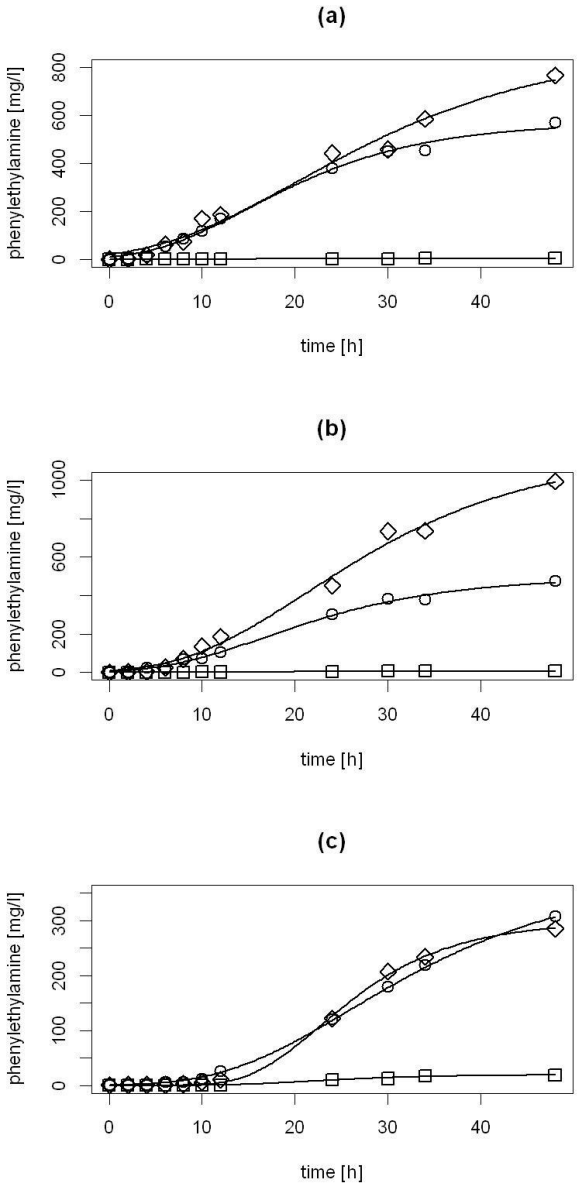


Fig. 1 Impact of pH of the environment on the asymptotic concentration and the maximum growth rate of phenylethylamine: (a) NaCl concentration 0 %; (b) NaCl concentration 2 %; (c) NaCl concentration 6 %. Markers at 11 time points denote average value of measurements (pH 5 circle, pH 6 diamond, pH 7 square). Corresponding Gompertz (or logistic) curves are drawn as well. Temperature is 30 °C, anaerobic environment.

B. Results for tyramine formation

Growing pattern was observed for all temperatures 6, 12 and 30 °C. Coefficient of determination R^2 realised in the range from 0.814 to 0.995. It should be noted that tyramine always gives substantially higher asymptotic concentration (parameter

A) than phenylethylamine. For instance, the maximal values of A parameter are 1,875 $mg \cdot l^{-1}$ (tyramine) and 1,134 $mg \cdot l^{-1}$ (phenylethylamine), respectively. The maximum asymptotic concentration was observed for temperature 30 °C, pH value 6, concentration of NaCl 1 % and anaerobic environment.

The effect of experimental conditions on value of parameter A

It seems logical that the lowest asymptotic concentration is given by temperature of 6 °C. For this temperature, values of parameter A decrease with increasing pH value. An interesting relationship involves aerobic/anaerobic environment, temperature 30 °C. The asymptotic concentration was higher under anaerobic conditions than in aerobic environment in the case of pH value equal to 7. In the case of pH of 5, aerobic environment provides higher concentration (and for the pH value of 6, the relation is unclear). Some other relationships could be formulated, however, there are more restricted to the levels of the remaining factors and therefore they are not mentioned. At a significance level of 5 % all the factors proved to be statistically significant. To be more precise, the hypothesis of A parameters equality was rejected in 96, 97, 86, and 74 % of all cases, respectively for the factor temperature, pH, NaCl concentration, and aerobic/anaerobic environment, respectively.

The effect of experimental conditions on value of parameter μ

The maximum growth rate is clearly influenced by temperature of the environment. It was observed that for increasing value of temperature (gradually 6, 12, and 30 °C) the value of parameter μ increases by at least an order (in average gradually 0.7, 7.5, and 222.5). At a significance level of 5 % all the factors proved to be statistically significant. The hypothesis of μ parameters equality was rejected in 100, 89, 77, and 59 % of all cases, respectively for the temperature, pH, NaCl concentration, and aerobic/anaerobic environment factor, respectively.

The effect of experimental conditions on value of parameter λ

The parameter λ takes its greatest value in the case of temperature 12 °C (in the range from 11.6 to 62.0). For the temperature of 30 °C the λ parameter values are substantially lower (between 0.3 and 13.9). Finally, for the lowest temperature 6 °C the lag time values took on negative values. The influence of all the factors tested was evaluated as statistically significant (at a significance level of 5 %). The hypothesis of λ parameters equality was rejected in 96, 92, 62, and 33 % of all cases, respectively for the factor temperature, pH, NaCl concentration, and aerobic/anaerobic environment, respectively.

The impact of pH of the environment and aerobic/anaerobic environment is illustrated in Fig. 2.

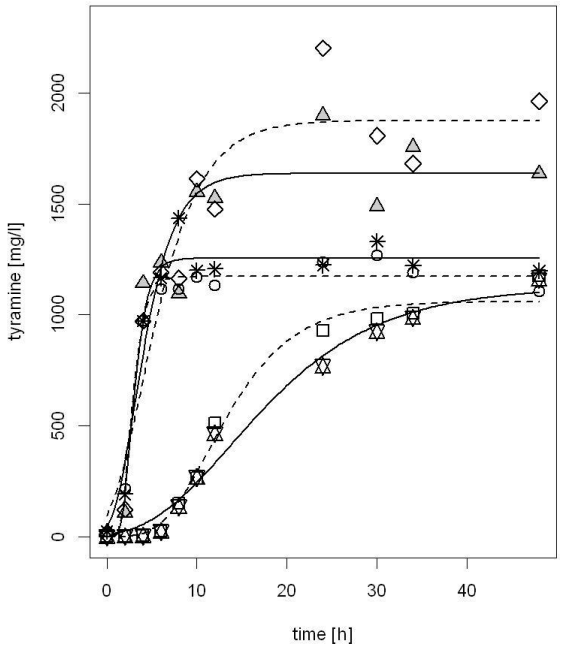


Fig. 2 Impact of pH of the environment and aerobic/anaerobic environment on the asymptotic concentration of tyramine. Markers at 11 time points denote average value of measurements (aerobic environment: pH 5 asterisk, pH 6 full triangle, pH 7 hexagram star; anaerobic environment: pH 5 circle, pH 6 diamond, pH 7 square). Corresponding Gompertz (or logistic) curves are drawn as well (aerobic environment: solid line; anaerobic environment: dashed line). Temperature is 30 °C, NaCl concentration is 1 %.

V. Conclusion

It was observed that *Enterococcus faecium* (DPE 002) produces high concentration only of two biogenic amines, namely phenylethylamine and tyramine. In the cases of other biogenic amines no concentration growing pattern is obvious (and the maximal mean concentration is less than 50 $mg \cdot l^{-1}$). In most cases, Gompertz or logistic model visually gave reasonably good fits of the data. In 87 % of the cases, the preferable model was the Gompertz than the logistic.

References

[1] L. Buňková, F. Buňka, E. Pollaková, T. Podešvová, and V. Dráb, "The effect of lactose, NaCl and an aero/anaerobic environment on the tyrosine decarboxylase activity of *Lactococcus lactis* subsp. *cremoris* and *Lactococcus lactis* subsp. *lactis*," *International Journal of Food Microbiology, 147(2),* pp. 112-119, May 2011.

[2] E. Dadáková, M. Křížek, and T. Pelikánová, "Determination of biogenic amines in foods using ultra-performance liquid chromatography (UPLC)," *Food Chemistry, 116(1),* pp. 365-370, Sep. 2009.

[3] L. Doudová, F. Buňka, J. Michálek, M. Sedlačík, and L. Buňková, "Risk analysis of tyramine concentration in food production," *11th international conference of numerical analysis and applied mathematics 2013: ICNAAM 2013,* vol. 1558, No. 1, pp. 1893-1896, AIP Publishing, Sep. 2013.

[4] A. Halász, Á. Baráth, L. Simon-Sarkadi, and W. Holzapfel, "Biogenic amines and their production by microorganisms in food," *Trends in Food Science and Technology, 5(2),* pp. 42-49, Feb. 1994.

[5] M. Kahm, G. Hasenbrink, H. Lichtenberg-Fraté, J. Ludwig, and M. Kschischo, "grofit: fitting biological growth curves with R," *Journal of Statistical Software, 33*(7), pp. 1-21, Feb. 2010.

[6] E. Lorencová, L. Buňková, P. Pleva, V. Dráb, V. Kubáň, and F. Buňka, "Selected factors influencing the ability of Bifidobacterium to form biogenic amines," *International Journal of Food Science and Technology, 49(5),* pp. 1302-1307, May 2014.

[7] P. Pleva, L. Buňková, A. Lauková, E. Lorencová, V. Kubáň, and F. Buňka, "Decarboxylation activity of enterococci isolated from rabbit meat and staphylococci isolated from trout intestines," *Veterinary Microbiology, 159(3),* pp. 438-442, Oct. 2012.

[8] M. H. Santos, "Biogenic amines: their importance in foods," *International Journal of Food Microbiology, 29(2),* pp. 213-231, Apr. 1996.

[9] B. Ten Brink, C. Damink, H. M. L. J. Joosten, and J. H. J. Huis In't Veld, "Occurrence and formation of biologically active amines in foods," *International Journal of Food Microbiology, 11(1),* pp. 73-84, Aug. 1990.

[10] C. P. Winsor, "The Gompertz curve as a growth curve," *Proceedings of the National Academy of Sciences of the United States of America, 18(1),* 1, Jan. 1932.

[11] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. Van't Riet, "Modeling of the bacterial growth curve," *Applied and Environmental Mikrobiology, 56(6),* pp. 1875-1881, June 1990.

Architecture of an Agents-Based Model For Pulmonary Tuberculosis

Moreno Luis Gabriel, Peña William, López Vargas Juan D.

Abstract- This text generally describes the architecture of an agent-based model for pulmonary tuberculosis in the zone of Usme (Bogotá, Colombia), product of a master's thesis in Information Sciences and Computing of District University "Francisco José de Caldas" [1]. First comes the introduction, then the tools and concepts that allow the simulation, then the model is exposed from a conceptual approach to technology-society relationship, and finally it is disclosed the authors' findings.

Keywords: Agents-Based Models, Geographic Automata System (GAS), Epidemiologic simulation, complex Systems.

I. INTRODUCTION

The technological edge thinks forward, looking for ways to understand it, study and intervene in the world over time towards a particular purpose. So it makes models that attempt to reproduce actual facts and allows a better understanding of the relationship of humans with them, appearing technologies with that purpose as Agent-Based Models and Geographic Automata System, among others [2].

But the future is thought from the uncertainty, because the world is complex and multiple, set of systems with the general characteristics of not to be reversible, not to be accurately predicted, to develop nonlinear interdependent relationships, and to have appearance of order (emergency), being necessary to go beyond the deterministic mathematical formal schemes [3, 4].

The bio-social systems, being complex, cannot be predicted with accuracy nor are reversible, so they have a constitution order from the simple to the complex, allowing the passage from chaos to order at the time called edge of chaos, where self-organization emerges, which is the coordination of the parts' behaviors of a system without any central power or external coercion that lead them. Jhon Stwart Kauffman formulated a hypothesis of a global connection between all parts of a physical system, that after a certain time, due to the energy accumulated between these, by inertia was that these associated themselves synergistically and generate patterns (self- adaptive systems) [3, 4].

Current knowledge have the task of describing the complexity of the world we inhabit, so in this case it is intended to represent the nonlinear and interdependent

relationships of a community in the middle of a tuberculosis epidemic, in order to predict the scenarios of this complex system [4, 5], explaining the construction of a Agent-Based Model for Tuberculosis in Usme's zone, and thinking about the origin and social horizon of technology.

II. TOOL AND BASIC CONCEPTS FOR THE MODEL

A. Geographic Automata System (GAS)

Between 1999 and 2001, Paul M. Torrens and Itzhak Benenson created Geographic Automata System (GAS) for modeling phenomena on real spaces, from individual computer entities within the system: agents. On the one hand, they modified the classical Cellular Automata (CA), which is a system in one, two or three dimensions, consisting of partitions called cells, which acquire ways to present: states, from a default set of them, ranging from relations with its neighboring cells, their neighborhood (Figure 1), through pre-established transition- state rules, in a sequence of moments called evolution system [6, 7].

And on the other hand, they placed those entities that change over time (A. Figure 2) on a Geographic Information System (GIS), which is a set of layers that describe the geographic characteristics of a place, from the general to the specific (Figure 2 B). Thus, agents are related to real spaces, preserving the neighborhood concept in his role for the change of state (Figure 1), but not only on generic objects such as cells, but sometimes on entities with own characteristics and mobility: specific sites , people, vehicles, etc. (C. Figure 2) [2, 6].

Fig 1. Types of neighborhood in cellular automata which allows the evolution of the system. Neighborhood A: Von Newman neighborhood, with octagonal neighbor cells. Neighborhood B: Moore neighborhood, with octagonal and diagonal neighbor cells [5].

A Geographic Automata System (G), is defined as the set of automatas (K), which vary from state (S) over time, through transition rules (Ts), developed at specific locations (L), from mobility rules (if any) of each one (ML), considering its neighborhood (N) at every time, called Tick, and the criteria of relationship they have with this (RN) (Figure 3) [7, 8]. Where, thanks to this set, It is able to see the interdependent relationships between the conditions of each area (population density, healthiness, etc.), and each agent location, nutrition, origin, economic status, etc.) [9].

ISBN: 978-1-61804-253-8

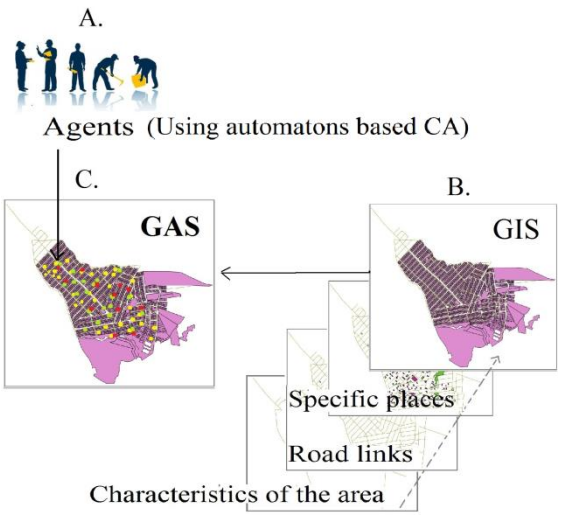


Fig 2. Integration of CA and GIS in the Constitution of GAS.

$$G \sim (\mathbf{K};\ \mathbf{S}, \mathbf{T_s};\ \mathbf{L}, \mathbf{M_L};\ \mathbf{N}, \mathbf{R_N})$$

Fig 3. Representation of a Geographic Automata System (GAS) [8].

First, the geographic information of Usme Central zone, was registered with ArgGis, application that allows to enter raster data type of a land, generating a GIS shapefiles (.shp) [10]. And then again, as Framework it was used Repast, which is a set of Java classes and methods linking agents with a GIS, allowing focusing on the modeling of the relevant attributes of the phenomenon, because this program already contains the multi-threaded programming, where in pseudo-parallel way for each Tick, transition rules are executed for all agents run, and also it provides a graphical interface of the simulation [11].

B. Usme zone

Usme is a local and administrative subdivision of Bogotá, which was incorporated into the city in 1990, because before that it was a town. So today some rural practices persist (agriculture, small farms, artisanal slaughtering), where according to figures consulted (data from 1999 to 2007 that varied in the study of late 2011, the same year it was supported the thesis on the model), 84.9% of its land (18,306.52 hectares) was rural and 15% urban, preserving some natural resources (now damaged by mining, urbanization and the tanning of leather) as water sources (21 in urban sector, 23 in rural sector, 11 rivers, 2 dams and 3 ponds.) [12].

Its population consisted of 51% female and 49% male, 34.8% of the population was under 15 years and 2.5% over 60, there being a high economic dependence, where on average 100 people were dependent of every 59, workers mostly with low levels of schooling and informal jobs. In addition, from 1995 to 2005, amidst the paramilitary phenomenon nationwide, this zone received 8.2% of the displaced population that came to Bogotá by the violence [12].

Thus, according to the Unsatisfied Basic Needs Indicator (NBI, by the Spanish acronym), which considers the shortages at home: a) housing with physical or structural defects, b) lack of basic services or deficiencies in drinking water and feces disposal, c) overcrowding (2 persons/5m2), d) high economic dependence (1 productive person / 3 or more dependents), e) truancy (at least one child between 7 and 11 do not regularly attend a school), and f) Misery, when the home has two or more of the above conditions, it was found that Usme was 9, 1% of homes in NBI, with 1% in misery. Fact also reflected in the 51% of the population below the "poverty line" indicator that arises from considering the minimum subsistence income for a person [13].

C. Tuberculosis Pulmonar (TB)

Illness from the bacteria Mycobacterium tuberculosis, native bovine and adapted to the human (zoonoses), which attacks places rich in mucous membranes (such as the lungs) and develops according to: the strength of the micro-organism to survive and be transmitted (virulence), the opposition that the body makes to the micro-organism (resistance), ease that people have to acquire and develop the disease (hypersensitivity) and the morphology of the affected tissues (genesis of the pattern infectious) being vulnerable to ultra-violet rays. It manifests in cough with coughing up phlegm or blood, evening sweating, fever, fatigue and unintentional weight loss.. Phases of development are: the attack or the arrival of bacillus in the body, then its logarithmic growth and the progressive activation of cells infected into other tissues, then there is an immunity, delayed as the body's inadequate response to disease, thus leading to the destruction of tissue and the transmission of new one [14].

Diagnosis is made by biochemical reaction of cultivated samples. To eradicate it is used treatment shortened supervised (DOTS), the antibiotic rifampicin, isoniazid, Pyrazinamide, and Ethambutol, over a period of 48 weeks, extending in the case of a relapse to 63, and whose success or failure (even fatal), depend on conditions of life of the infected as food, hygiene, wholesomeness in the habitat and permanence in the treatment, since deaths from TB are commonly associated with poverty and undernutrition [15].

III. Agent-Based Model for Tuberculosis in Usme

A. Simulation's Time and Space

In architecture model time and space simulation, as the first condition for the interdependence between agents is carried out in the middle of tuberculosis outbreak as a complex system were established. On the one hand *TimeProject* whose *TimeProject* constructor class is created generated internal time model, allowing the passage of the method *step()*. And moreover the *Global* class, entering shape files and generates graphical output interface was created, also it determines the amount of Tick equivalent to every hour (1 hour = 1 Tick) , the number of initial agents in each state, that the epidemiological model of pulmonary tuberculosis (SIR model) are: Susceptible (which could be infected), infected (disease carrier) and Recovered (who overcomes the infection), and the amount of each type of agent , which under Usme socioeconomic conditions are: *Housewife*, *Worker*, *Student*, *Deplaced* and *Homeless*, which enter as parameters

via *setValuesFromParameters()* method of that class [1] (A. Figure 4).

Therefore, States and agents that is different in the middle of a space and a time shared, determine that the Organization of the information in a complex system should be a dynamic structure that contains the particularity of the elements and their no-lineal relationships, and don't simply a collection of attributes.

B. Contexts

Contexts are joint that grouped the agents of a system and the relationships between them. Thus, *MainContext* class is the generic context that through the *build()* method, load the global timeline, originates and controls the contexts of spaces and people, and is the time and the particular space of these agents. Subsequently determined the context of each place: Home (*HomeContext*), workstation (*WorkPlaceContext*), study location (*StudyPlaceContext*) and entertainment venue (*EntertainmentPlaceContext*), dynamically created by reflection (in the execution of the program), through the method *createSubContext()* in the class *CityContext*, referencing their positions from the tract (class *RoadContext*), the intersections of these and the boundaries between places (class *JunctionContext*). In the same way, with the class *PersonContext* was created the possibility of dynamically generating the contexts of agents -sets of sites and other agents-: *HousewifeContext*, *WorkerContext*, *StudentContext*, *DeplacedContext*, *HomellessContext* (B. Figure 4).

The location of the agents was determined by the rules of movement (ML) of each, using for this parametric class *AgentContext* and *getGeography()* method. And the location of the sites was performed using the same method name *PlaceContext* parametric class (C. Figure 4).

Due to the needed for information on the TB outbreak (major focus of infection, more infected population trend of spread of the disease, etc.), which would act on this, the dynamic generation of contexts allowed the development of the non-linear relationships between the different agents in a period of time, loading and processing the increased volume of information in the model thanks to the architecture of the program, through reflection, extended the possibilities of development of the system.

C. Places

Parametric classes system can evolve on the geographical conditions of Usme, which is very important for the relationship between the demographic conditions of a region and the spread of an epidemic (TB in this case), where the different types of places and their characteristics were determined and verified with the abstract class *Place*, with the attribute *listPerson* created an arrangement for certain amount of agents inside, regardless of their type, verify with the method *getOvercrowdingCondition()* the overpopulation (boolean data), on the basis of the capacity of the place (method *getPeople Capacity()*), and the number of people in it (method *getAmountPerson()*) [1].

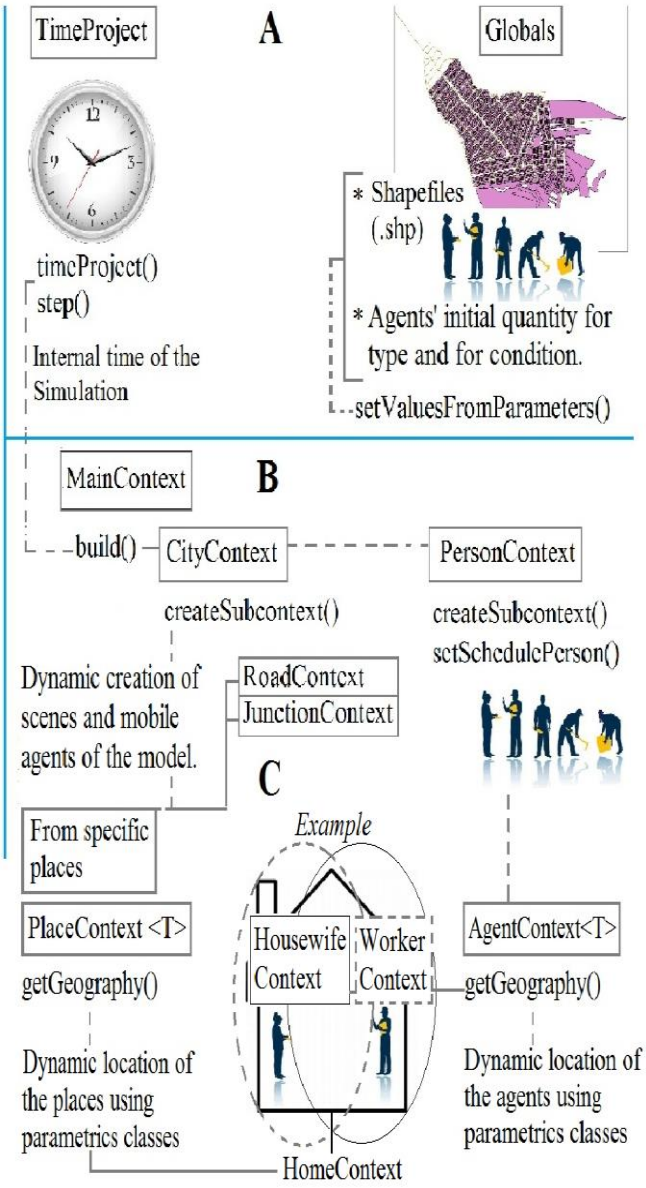


Fig 4. Time and space model, and training contexts through parametric classes.

The specific locations and attributes were defined with *HomePlace* classes (households), *WorkPlace* (work sites), *StudyPlace* (study sites), *EntertainmentPlace* (entertainment venues) and *Route* (tracks), which inherit of the class *Place* attributes as the maximum area place (*MaxArea...*), high capacity (*MaxAmount...*), many people (*amountPersons*), capacity (*peopleCapacity*), use, management, geometry, number of people per state (*amountSusceptible*, *amountInfected*, *amountRecovered*), among others.

Calculation of amounts by State was performed on the method *step()* -different class *TimeProject step()* method, and later exposed the *Person* class *step()* method-, and the distribution of agents by States was random from classes of each place: *Work Place*, *Study Place*, *Home Place*, *Entertainment Place*, trough the method *getRandomStudyPlace*, *getRandomHomePlace*, *getRandomWorkPlace*, and *getRandom EntertainmentPlace*, respectfully (A. Figure 5).

The description of Usme geography through Shape files and their inclusion in the system, allowed to generate

scenarios appropriate for studying the TB epidemic in the actual conditions of this place, as a requirement of a model oriented to care for the life of the population, through the generation and the study of patterns in the spread of this disease.

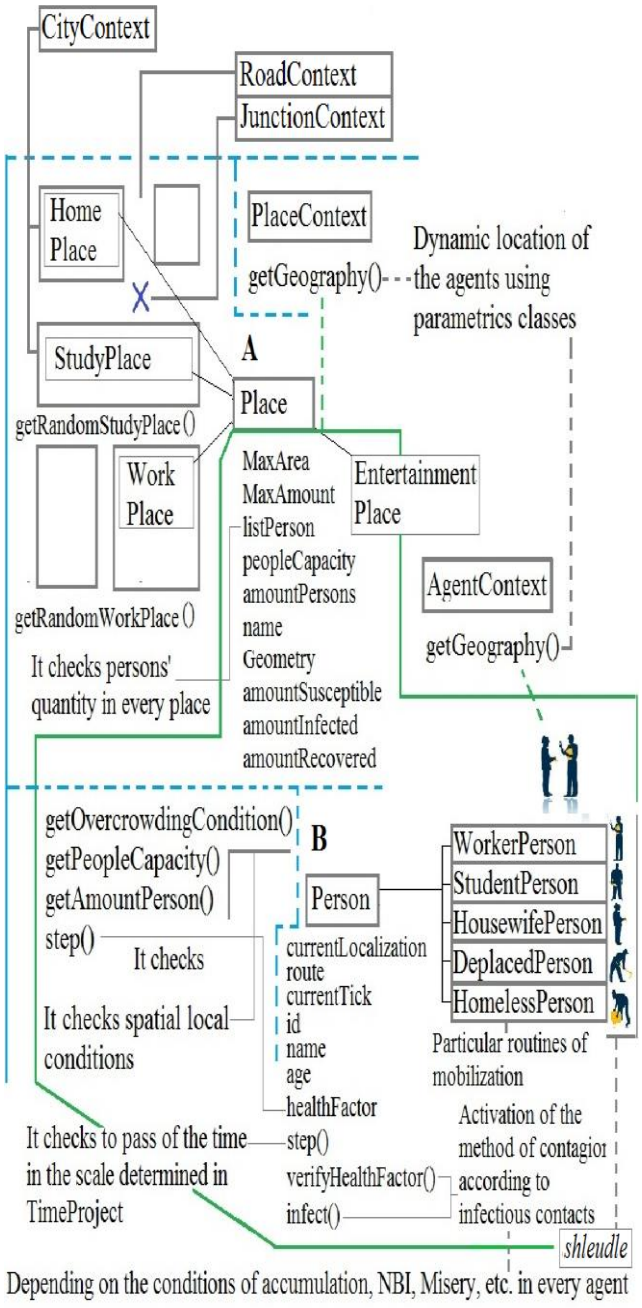


Fig. 5. Architecture of Sites Agents Model.

D. Agents

It determined each agent-specific property so that they can evolve relationship tailored to the reality of the TB outbreak: daily places of movement, origin in the case of displaced persons, etc. Why the particular agents were created 'Worker', 'Housewife', 'Student', 'Deplaced for violence' and 'Street people', through classes, *Worker*, *Housewife*, *Student*, *Deplace* and *Homeless*, respectively.

Previous classes inherit from the Person class attributes of localization in each place (*currentLocalization*), routes of movement (*route*), location at every moment of the system (*currentTick*), identification (*id*), employment status, medical service, stratum, previous infection by HIV and TB, type of person, State of health (*healthFactor*) and number of health (*numberHealthFactor*), which is verified by the *verifyHealthFactor()* method According to the infections activate method *infect()*, which passes the person State Susceptible to infected (first passive, then contagious).

Although it must be clarified that home and the places of development of each agent are attributes in each particular class (*Worker*, *Student*, *Housewife*, *Homeless*, *Deplaced* etc.), and is thus allowing you to determine routines on different types of agent through parameters that are entered as *Schedule* (specific cases) in the classes of every kind of person (B. Figure 5).

E. Rules of Transition

Rules of transition between the States of SIR epidemiological model are identical for all agents (TABLE I.): susceptible to passive infected, and hence a contagious agent, then be recovered, however the times varied considering multiple living conditions of Usme (TABLE II.), where the method *infect()* of class *Person* activated particular rules contained in the classes of each type of agent making it easier in some cases to the acquisition of disease, and likewise, hampering its recovery (average increase of vulnerability, TABLE II) [1].

TABLE I: TRANSITION RULES.

State change	Susceptible to Infected liability	Infected liability. From Infected Contagious	Infected to Recovered
Time (1 Tick = actual 1 hour).	12 Tick	96 Tick	Tick 4032 (24 weeks): Assuming completion of treatment of 48 weeks.

According to the conditions of life in each agent, for example, the time 12 Tick in the vicinity of a contagious, necessary to move from Susceptible to passive infected, agent reduced by the average decrease in 1/8, 1/6, or 1/4 of the number healht Factor attribute, as there was an average delay of 1/8, 1/6, or 1/4 in raising that attribute according to the conditions of table 2, to move from infected to retrieved, for a population chosen randomly according to the percentage of population affected, with the following conditions: 1 in every 4 people in overcrowded housing was inadequate, 1 in 3 people in destitution was in misery by NBI, 1 in 4 people with poor nutrition in the NBI1 in 3 people with chronic malnutrition was on the same indicator, as gaps in living conditions have problems shared, setting such conditions through schedule in each type of agent [9].

TABLE II.: IMPACT ON VULNERABILITY FACTORS ATTRIBUTE PERCENTAGE WITH POPULATION AFFECTED *NUMBERHEALHTFACTOR*.

Vulnerability Factor	**Average increase in vulnerability. Attribute**	**% Affected Population** [1]

[1] Estimated figures weighting of survey data quality of life and social diagnosis of the town, considering older vulnerabilities for IDPs and resident street [9, 10].

	numberHealhtFactor () <= 0.2 = infected. Scale of 0-1.	
Poor nutrition	1/8	30%
Chronic malnutrition	1/6	15%
Overcrowding	1/4	5%
Housing inadequate	1/6	2%
Households with NBI (1 Home = Average of 4 People)	1/8	9.1% households = 36.4% population
Misery as NBI	¼	1% of households = 4% population
Homelessness	¼	6%

F. A CASE

With a total amount of 500 agents, 20% love House, workers 25%, 25%, 15% and 15% displaced students of street, 5% initial of infected, recovered 20% and 75% susceptible, were obtained data from Table III.

TABLE III: AGENT TYPES OF INFECTION BY TICK.

Agents	Infected% (of the total) for Tick. T= 1	T= 200	T= 528	T= 4032
Housewife	1%	1.2%	3.8%	0.2%
Worker	1%	1.8%	4.2%	0.4%
Student	1%	1.4%	3.4%	0.2%
Displaced	1%	2%	3.6%	0.8%
Homeless	1%	1.6%	4.2%	0.6%
TOTAL INFECTION	5%	8%	19.2%	2.2%

The resulting data is inferred that infection levels rise first in patients with immune deficiencies related to bad nutrition, poverty, and inadequate conditions of habitat (housing overcrowding or free services: NBI), nothing that although the highest proportion of agents corresponded to a type (workers and students), harder those infected in that eradicating the disease, they are those who described under very poor conditions of existence (Street and displaced inhabitants), which, if they had not taken into account would have caused a uniform tendency in the simulation behavior, without allowing to see the auto-organising patterns of the disease from the actual conditions to be able to act on this.

IV. CONCLUSIONS

1) The dynamic construction of contexts from parametric classes (T:Microsoft.VisualStudio.Test Tools. execution. agent context and Place Context) in the architecture of the model allowed a greater volume of information loaded and processed, reducing the lines of code and optimizing the performance of the machine, but projecting a greater precision of the simulation to the real conditions of the phenomenon in Usme (special agents(: street people, displaced persons, and specific conditions: NBI, nutrition, overcrowding) at all times, making this program as a support tool that could be used by local health authorities.

2) The creation of Time Project and Global classes with their respective methods, and the union coherent types of agent to these, incorporating in a single set all elements of the simulation, is the architecture of the system as a Framework for the creation of models of epidemics-targeting GAS, regardless specific geography in which develops and simulated disease now that the creation of common space and time, and the possibility of agents to act is through these by means of dynamic contexts, create the basic conditions for any model can develop.

3) The model is created for the study of infection and the spread of tuberculosis taking into account the social dynamics, adopt to different fields and learn behaviors and theoretical but reliable of a possible epidemic statistics without ethical implications, and allowing the application of preventive methods and control large scale.

REFERENCES

[1] Moreno Sandoval, L. G. Epidemiological model of the tuberculosis based on agents in Usme, Bogota, Colombia. Thesis of mastery in Sciences of the Information and the Communication. University Distrital " Francisco Jose of Caldas ". June, 2007 - March, 2012.

[2] Benenson I., Torrens P. M. Geographic Automata Systems: A New Paradigm for Integrating GIS and Geographic Simulation. In Proceedings of the AGILE 2003 Conference on Geographic Information Science, Pages 367-384 (Lyon, France). April 24th- 26th, 2003.

[3] Lucas C. Self-Organization. In Magazine Thinking the complexity, 3(6): 5-15, 2009.

[4] S. M. Manson., S. Sun., D. Bonsal, "Agent-Based Modeling and Complexity" in *Agent-Based Models of Geographical Systems* A.J. Heppenstall, A.T. Crooks, L.M. See, M. Batty, Eds. Dordrecht, Heidelberg, London, New York, Springer, 2012, ch, 4, pp. 125-140.

[5] Perez Martínez A. Stuart Kauffman's work. The problem of the complex order and his philosophical implications. In Magazine Thinking the complexity, 3(6): 21-38, 2009.

[6] Benenson I., Torrens P. M. Geographic Automata Systems. In International Journal of Geographical Information Science. 19 (4): 385–412, April 2005.

[7] Kari J. Cellular Automata. Technical Report. University of Turku, Finland, 2013.

[8] I. Benenson, V. Kharbash. Geographic Automata Systems: From The Paradigm to the Urban Modeling Software. In: Proceedings of the 8th International Conference on Geocomputation University of Michigan. United States of America. July 31 to August 3, 2005.

[9] M. Parvin, "Agent-based Modeling of Urban Phenomena in GIS" Capstone Project, thesis submitted in fulfillment of Masters in Urban Spatial Analytics, University of Pennsylvania, CA, 2007

[10] Autonomous university of Madrid. Manual (basic level) for the production of maps with ArcGis. December, 2011.

[11] Malleson N. Repast Simphony Tutorial. Technical Report. May 7, 2008.

[12] Secretariat of estate of the District. Crossing Usme 2004. In: physical Physical and socioeconomic Diagnosis of the localities of Bogota, D.C. Major mayoralty of Bogota. Bogota DC, Colombia. 2004.

[13] Secretariat of Health of the District. The health and the quality of life. Locality 5 - Usme. Major mayoralty of Bogota DC. 2009.

[14] Dr. S. Invertrz. Pathogeny of the tuberculosis. University of Buenos Aires. Buenos Aires, Argentina. 2008.

[15] Dr. Ospina S. The tuberculosis, a historical - epidemiological perspective. Magazine Infectio. Vol.5 no. 4 pp. 241-250, Bogota October - December, 2001.

First Author: Luis Gabriel Moreno, Systems Engineer Liberators University, Masters in Computer Science and Master of Digital Marketing externship Colombia University and MBA candidate at University of Granada Spain Docotroando CEO Thinking Colombia

Second Author: William Peña, Systems engineers and Masters in Computer Science professor at Colombia Uniminuto

.

Third author: Juan Diego López Vargas, Telecommunications Engineer Santo Tomas University of Colombia, specialist, and doctoral magister Integration of Information Technology in Organizations Livelihood at the Polytechnic University of Valencia Spain. Consultant and Director of Special Projects and Research at Colombia University graduate Thinking Ecci Colombia.

Investigation and Analysis of functional performance between Tibetan and Han university students in Gansu

Bai Jingya, He Ye*, Hai Xiangjun, He Jinquan, Wang Yutang, Wang Zijiang

***Abstract*—**The research purpose of this paper is to compare and study the functional performance of Gansu Tibetan university students and Gansu Han university students. Cross-sectional random cluster sampling method was adopted to survey the body height, body weight, BMI, vital capacity, body weight index, step index, body bending on seat, 50m sprint etc. of the students in Northwest University of Nationalities in 2011, in which body height, body weight and BMI index are used to evaluate physical shape index; BMI index classification standard is used to evaluate nutrition status; vital capacity, body weight index and step index are used to evaluate physical function; body bending on seat, 50m sprint indexes are used to evaluate physical quality. Results showed the physical shape indexes of university students of both nationalities were within normal range. Compared with Han university students, Tibetan university students showed better situation in nutrition with slow growing development speed, and male showed poorer and female showed better conditions in body enrichment and development degree. In the development of physical function, university students of both nationalities showed poor condition in lung function and need to be further improved, while Tibetan university students were poorer than Han university students. The cardiac function of university students of both nationalities performed in good condition, and Tibetan male showed better condition than Han university students. In the development of physical quality, the flexibility quality of both Han and Tibetan university students were good but Tibetan males were better than Han males and Han females were better than Tibetan females. University students of both nationalities performed poor in speed quality (especially female). Comparison in functional performance with Tibetan university students in other regions showed Tibetan university students of Gansu Province were poorer in physical shape index but better in function index.

***Keywords*—**Gansu, Tibetan, Han, university students, functional performance .

I. INTRODUCTION

COLLEGE time is the last period of puberty development and the key period for the development of psychological and physical health and all physical qualities of university students. The growing development level and health condition in this period will produce significant impact on their physical quality, psychological quality, employment, occupation selection and future daily life[1]. University students are national pillars and their physique health condition not only personally matters but also concerns the "healthy quality" of the whole nation and even the whole country. Practices and researches of the countries all over the world indicate that it is effective to prevent "health problem" through the improvement of physique level. Therefore, exploring into the physical and psychological development law f university students and how to improve the physique of university students have been the focus about health of all countries and the hot spot of research on physique[2].

The research on physique of university students has been a hot issue of health education, public health and sports medicine fields in recent years. In order to better push forward the implementation of university student health promotion plan, naturally it becomes more important about how to survey and collect basic data of physique health condition, physical activity pattern, daily life physical exercise behavior, health risk behavior and other relevant areas, and how to make combined qualitative and quantitative measurement and evaluation. At current, most researched are focused on the measurement and evaluation of health-related physical fitness and physical activities of university students, and some researches, such as the research on obesity of university students, also combined metabolic-related index and hematology index tests [3]. Although researches on physique have made remarkable achievements with years' efforts, most researches were targeted on children, teenagers and adults instead of this special group-university students. Whether the content, method and evaluation system of the current physique health test can reflect the current growing and developing status of university students in an accurate and comprehensive manner and work as the basis to guide scientific fitness of university students are still greatly controversial in implementation and yet to be proved and perfected by physique workers [4].

Influenced by innate factors such as heredity and postnatal factors such as environment, nutrition, health, sports exercise and so on, university students of different nationalities living in different regions formed different characteristics of physique. Tibetan people are one of the major minorities of China, and the functional performance of Tibetan university students has attracted much attention in recent years. Wang Cheng made a research on the functional performance of Tibetan and Han

This work was supported by the Fundamental Research Funds for the Central Universities(Grant No.31920140068) and by Chengguanqu scientific and technical program (2014-6-3)

Corresponding author, e-mail: leidou0315@qq.com

university students and indicated that Tibetan students in some universities of China had better condition in physical developing level, overall vital capacity level, lower limb power etc. than Han university students[5]. She Jingfang made a research on the functional performance of Tibetan and Han university students inland and indicated that inland Tibetan university students had better condition in respiratory function and aerobic stamina level than Han university students; while they had poorer condition in upper limb power, explosive force, sensitivity than Han university students [3]. Ding Min etc. made a research on the functional performance of Tibetan and Han university students in Tibetan area and indicated that male Tibetan university students were at the similar level with male Han university students in physical development and the overall vital capacity level of Tibetan university students were lower than that of Han university students on average, while the physiological function and endurance level of cardiovascular system were significantly better than that of Han university students [4] .

Gannan Tibetan Autonomous Prefecture and Tianzhu Tibetan Autonomous County are the two major Tibetan regions of Gansu Province. In recent years, significant social, economic and cultural changes occur in this two Tibetan regions; the living standard of people is greatly increased, and survival and development situation is greatly improved, laying solid substantial basis for the growing and developing of Tibetan university students and allowing considerable and sustainable development of the physical development of Tibetan university students[6-7]. In [8], it explored the relationship between family average income (FAI; an index of socio-economic status) and body mass index (BMI; a widely used, inexpensive indicator of weight status) above the healthy weight range in a region of Mainland China. The proportion of adults with BMI above the healthy weight range was positively related to having a higher socio-economic status (indexed by FAI) in a regional Chinese population. But research materials about the functional performance of Tibetan university students of Gansu Province are absent at current.

This paper is to make a research on the functional performance and characteristics of Tibetan university students of Gansu Province and compare them with Han university students of Gansu Province to explore the reasons causing difference of physique health so as to further improve the physique health condition of Tibetan university students of Gansu Province and provide reference and basis for physical education reform of colleges and universities.

II. METHODOLOGY

A. Sample selection

In October 2011, stratified random cluster sampling method was adopted to make investigation on Tibetan and Han students of Gansu Province of Northwest University of Nationalities. Tibetan university students were required to be those who live in Tainzhu County and Gannan Prefecture in long-term basis and Han university students in Gansu, and all of them must be between 18～23 years old and healthy without disability and chronic diseases. Tibetan university students were sampled separately from rural and urban areas, and Han university students were sampled with equivalent number from upper, middle and lower social and economic areas throughout Gansu Province. 1052 students in total were as samples, including 344 Tibetan university students (male, 174; female 17) and 708 Han university students (male, 283; female, 425).

B. Measurement Methods

According to the requirements of Physique Health Test Standard of University Students, professionals tested the body height, body weight, vital capacity, standing long jump, step index, body bending on seat of university students, and all quality control indexes on spot conformed to detailed requirements to ensure the scientific test and accurate test results.

Height: standing height and sitting height meter was used as the instrument to measure with marking off in centimeters. Subjects were asked to take off his shoes and hats, stood on the floor at attention, hang naturally down hands, bring heels closer together, stands apart toes to about 45 degrees apart, as well as the three points of heels, hips and shoulders were asked to close up the column, the trunk is were asked to naturally hold themselves upright, the two eyes were asked to level with the front. Moreover, the examiner was asked to stand in the right side of the measuring board, and gently move the slide until reach the head point. The measurement error shall not exceed 0.5 cm.

Weight: the instrument of the leverage scale was used for the weight measurement in kilograms. The subjects were asked to deplete the urine and underwear, and were asked to wear underwear, stand over in the middle of the weighing platform with bared feet. As well as the hands did not touch other objects, and the counterweight was adjusted until the balance of leverage and the data was recorded to the smallest scale. And the test error is not more than 0.1 kilograms.

Vital capacity test: the subjects stand facing the instrument and hold the blow inlet nozzle to take twice deeper breath than usual, then breathe in deeply, and hold the breath to exhale at the nozzle to the greatest extent. After blowing, the figure showed on the liquid crystal screen is the milliliter value of vital capacity. Each shall be tested every 15 seconds for three times in total. Values of all three tests will be recorded and the largest one will be taken as the test result.

Step index test: before the test, the pulse in peaceful moment should be tested, then tested will do some mild preparation activities to mainly exercise lower limb joints. The subjects stand upright in front of footsteps (40cm-height of footstep for male and 35cm-height footstep for female) to step on and off the horse block according to the prompt tone sent by the metronome. When the first sound of the metronome is made, one foot steps on the block; when the second sound is made, the other foot steps on and two legs stand straight; when the third sound is made, the first foot steps off; when the fourth sound is made, the

other foot steps off. After repeating continuously for 3 minutes, the subjects immediately go back to chairs in silence. Record the pulse 1-1.5 min, 2-2.5min and 3-3.5min respectively after the exercise. If the subjects can not finish the stepping-on and stepping-off exercise according to the tempo of the metronome in this three repeats or fail to insist, he/she should stop at once; record the lasting time of the exercise and also record the pulse rate for three times in the same way. Then step index can be calculated with the following formula:

Evaluation index = Lasting time of stepping-on and stepping-off exercise (sec.) ×100/ 2× (the sum of pulse rate measured of the three times)

Body bending on seat test: in face of the instrument, the subjects sit on the cushion with double legs straight forwards, two heels close together pedaling on the damper of the tester, tiptoes naturally separate, both hands close together, palms downwards in full stretch, knee joints stretched and the trunk bent forwards; and then push the cursor slide forward with the middle fingertips of both hands to the greatest extent. Record the test value in two tests and take the larger one by centimeter.

50-meter sprint test: the subjects are tested in group containing at least 2 subjects. Standing start, the subjects start to run upon hearing the command “Run”. The starter should swing the commanding flag while giving the command. The timist starts to keep time as the flag moves and stop keeping the time when the trunk of subject reaches to the vertical surface of the finishing line. Record the results by second.

C. Quality Control

The measurement crews are strictly trained. The measuring instruments are calibrated before survey. The measuring time is limited within 8:00am - 11:30am every day. The testing room shall be well ventilated. And subjects are required to wear short pants and blouse. Strict quality control measures are required in survey location selection, test time, test methods, instrument type etc..

D. Data Processing

Perform statistical processing on all data through Excel2003 and SPSS19.0 software.

III. RESULTS

A. Physical shape index

Seen from Table 1, the body height, body weight, BMI indexes of both Tibetan and Han university students are within normal range. Average body height of both Han males and females is higher than that of Tibetan male and female; the body weight of Han males on average is better than that of Tibetan males; as for BMI index (reflecting the sturdy degree of figure), Han males are better than Tibetan males and Tibetan females are better than Han females, which are all of statistical significance.

Seen from Table 2, both Tibetan and Han male students with normal body height take up the highest proportion, and the next is overweight, underweight and then obesity. both Tibetan and Han female students with normal body height take up highest proportion, and for Tibetan female, the next is overweight and underweight, then for Han female, the next is underweight, overweight and then obesity. Tibetan males and females with normal body weight take up a larger proportion respectively than Han males and females, and Han females underweight take up a larger proportion than Tibetan females, which are all of statistical significance.

B. Physical function index

Seen from Table 1, the overall vital capacity of Han males is higher that of Tibetan males and the vital capacity body weight index of Han females is higher than that of Tibetan females, which all of statistical significance. Besides, the vital capacity body weight of Tibetan and Han university students are at the qualification level of Physique Health Standard of University Students.

Seen from the step index which reflects physical cardiac function capability, both Tibetan and Han university students are up to excellence level of Physique Health Standard of University Students and Tibetan male shows better performance than Han males, which are of statistical significance.

C. Physical quality index

Seen from Table 1, the body bending on seat index value of both Tibetan and Han university students are up to excellence level of Physique Health Standard of University Students and Tibetan males show better performance than Han males, while Han females show better performance than Tibetan females, which are all of statistical significance.

The average results of 50-meter sprint of both Tibetan and Han males are up to pass level of Physique Health Standard of University Students and the average results of 50-meter sprint of both Tibetan and Han females are all at failure level of Physique Health Standard of University Students.

IV. DISCUSSION

Body height is an effective index to describe physical longitudinal development and also one of the most direct indexes reflecting physical shape. Body weight is an effective index to describe physical transverse development reflecting the comprehensive development status of physical bones, muscle, subcutaneous fat and internal organs. BMI is a comprehensive shape index reflecting the substantial degree of human bodies and greatly significant to evaluate the physical proportion, growing development and nutrition condition of university students[6].

Table 1 analysis of Physical shapes, Physical function and physical qualities indexes between Tibetan and Han university

students in Gansu in 2011（$\bar{x}$±S）

nationality	gender	sample	statistic	Physical shapes index			Physical function index			physical qualities index	
				body height /cm	body weight /kg	BMI(kg/m2)	vital capacity/mL	vital capacity body weight index(mL·kg-1)	Step index	Body bending on seat /cm	50m/s
Tibetan	male	174		170.82±5.45	62.30±7.92	21.35±2.49	3643.61±593.65	58.26±8.58	26.38±10.15	15.67±6.02	7.63±1.40
Han		283		172.55±7.11	65.36±11.24	21.9±3.13	3801.03±659.97	58.21±8.72	53.03±8.13	14.03±6.53	7.61±1.35
			t value	2.931	3.406		2.571	-0.60	-3.882	-2.675	-0.138
			p value	0.004	0.001		0.01	0.952	0.000	0.008	0.890
Tibetan	female	140		158.79±5.27	53.42±6.69	21.17±2.31	2454.96±474.74	45.76±8.45	50.87±7.11	15.86±5.66	9.17±0.52
Han		425		160.14±5.05	52.70±6.28	20.56±2.33	2518.13±416.79	47.55±7.58	51.15±8.29	17.62±6.55	9.15±0.81
			p value	2.712	-1.167		1.501	2.356	0.352	3.062	-0.321
			p vales	0.007	0.244		0.134	0.019	0.725	0.002	0.748

Table 2 analysis of nurture between Tibetan and Han university students in Gansu in 2011 (n(%)^2)

nationality	gender	Samples	BMI（kg/m^2）			
			<18.5	18.5～23.9	24～27.9	≧28
Tibetan	male	174	12（6.9）	139（79.89）	19（10.92）	4（2.3）
Han		283	33（11.66）	192（67.84）	45（15.90）	13（4.59）
Tibetan	female	140	11（7.86）	114（81.43）	15（10.71）	0（0）
Han		425	87（20.47）	304（71.53）	31（7.29）	3（0.71）

Comparison of Gansu Tibetan university students with Gansu Han university students in 2011 showed, in physical shape development, the growing development speed of Tibetan university students was slow, while in physical substantial degree and development degree, males were poorer and females were better; in nutrition Tibetan university students showed better condition than Han university students, which indicated that the physical shape index development condition of Tibetan and Han university students (excluding Han female) was in compliance with their nutrition condition; Han females pay more attention to the demand on shape and keep fit through inappropriate methods such as going on diet, which results in serious malnutrition.

During decades of years' development of physique researches, the research contents are also continuously changing and health-related physical fitness research has dominated the physique research content. The tending of physique research to promote health service is more and more obvious and continues to convey such message to people: good physique means good health and the ability to engage in labor activities safely and prevent various diseases caused by lack of exercise. The physique research direction and priorities in constant correction eventually combined physique and health closely to make is a rigorous and practical science.

In this paper, the physical function is measured with vital capacity and step test indexes, among which, vital capacity is an index reflecting respiratory function and closely related to body height, body weight, and chest circumference. Generally university students have stepped into adult stage when the influence of body height is significantly weakened while the body weight becomes the major physique growth index that influences the level of vital capacity. In general, adults with heavier body weight also has higher level of vital capacity, and most of the vital capacity is used to overcome physiological compensatory expressed by physical load. Therefore, calibration with vital capacity body weight index can better reflect the condition of individual or group lung function. Step test is a simple and practical quantitative load test, and the step index calculated from test results is a comprehensive index combining load time, exercise reaction and heart rate recovery speed after load, reflecting the level of physical cardiovascular system function. The higher the step index is, the better the function of physical cardiovascular system is. The research results of this paper indicate, in physical function development, university students of both nationalities show poor condition in lung function and need to be improved, among which, Tibetan university students were poorer than Han university students. The cardiac function of university students of both nationalities were in good condition but Tibetan male was better than Han university students. Tibetan university students selected in this paper are from two Tibetan regions of Gansu, where the altitude is high, the weather is freezing, sunshine radiation is strong and the air is thin, which make great impact on physical physiological function. In [7], it indicated, for 90% of people

who live on plateau, their breath frequency slowed down, breath amplitude went shallow, and all physical ventilation indexes declined under the environment with normal pressure and abundant oxygen, while cardiovascular system had the capabilities to store and convey nutrient substances, oxygen and metabolic products etc., which objectively promoted the physiological function of the cardiovascular system. For Tibetan university students who study on plains and adapt to plain conditions, their physiological function showed a conservative declining trend, which resulted in that the overall average vital capacity level of Tibetan university students was lower that of Han university students, while their physiological function of the cardiovascular system was significantly better than that of Han university students.

In this paper, body bending on seat index and 50-meter sprint index are selected to measure physical quality index. Body bending on seat is to measure the potential motion extent of joints such as trunk, waist and hip etc. of university students in the rest of rest, reflecting the extensibility and flexibility of joints, ligaments and muscles at these positions as well as the development level of physical pliable quality of university students. 50-meter sprint is to test the development level of the speed, sensitive quality and the flexibility of nervous system of university students. In the development of physical quality, both Han and Tibetan university students showed better condition in pliable quality, of which, Tibetan male was better than Han males while Han female was better than Tibetan female; and the university students of both nationalities showed poorer speed quality (especially female), which is directly related to their lack of sports exercise.

From the comparison (+ test) of the functional performance of Tibetan university students in other regions, we know Tibetan university students of Gansu Province showed poorer condition in physical shape index than Tibetan university students of other regions and better in function index than the average level of students in some universities all over the country. The vital capacity index of Gansu male university students was higher than that of Tibetan male university students in Tibetan areas, while the step index and physical quality index were lower than them.

REFERENCES

[1] BAO Xue-ming,JI Cheng-ye,YIN Xiao-jian. (2008). Analysis on the trends of the chinese university students' physical change from 1985 to 2005. Modern Preventive Medicine. 3364-3367

[2] WANG Cheng, LI Yong-jie,SHI Ru-lin. (2008)A Comparative Study of Physical Conditions of College Students from Han and Tibetan Nationalities. Journal of Beijing Sport University. 7(7):946-953

[3] SHE Jing-fang. (2009).Comparative anglysis of constitution and health of college students of inland Tibet and Han nationality. Liaoning Sport Science and Technology. 48-49

[4] Ding Min, Zhao Fengcang, Lu Juan. (2009). Analysis of physical examination results between Tibetan and Han freshmen in Tibet University for Nationalities, Chinese Journal of School Health,754-755.

[5] SHAO Yan. (2009).An analysis of the influence and efficacy on students from student physical health standard implemented in Jiangsu province. Journal of Zhoukou Normal University. 145-147.

[6] Carter JEL , Heath BH . Somatotyping Development and Application.London: Cambridge University Press，1990

[7] Jiang C, Liu F, Luo Y, Li P, Chen J, Xu G, Wang Y, Li X, Huang J, Gao Y (2012) Gene expression profiling of high altitude polycythemia in Han Chinese migrating to the Qinghai-Tibetan plateau. Molecular Medicine Reports .5(1):287-293.

[8] Xu F,Yin XM,Zhang M, (2005).Family average income and bodymass index above the healthy weight range among urban and ruralresidents in regional Mainland China. Public Health Nursing . 47-51

Study of Wastewater Treatment Plant

S. Al Jlil* and M. Sajid

***Abstract*—** The study was carried to evaluate the performance of nanofiltration (NF) and reverse osmosis (RO) technology for reducing the total salt concentration from waste water. The nanofiltration proved very effective in removing the polyvalent cations and anions such as SO_4^{2-} where the removal efficiency was 97.22 %. In is well known that the RO water treatment process removes all the cations and anions from waste water or brine or sea water especially removing the monovalent ions such as Cl^- where the percentage removing efficiency was 94.4%. The performance efficiency of RO and NF water treatment processes declined significantly during the first 3-years of operation due to fouling and biofouling of the membrane. These significant findings provided concrete clue for the important issue to replace or reject the existing water use methods. The research further highlighted the necessary to replace the NF and RO membranes used in these two water treatment techniques..

***Keywords*—** Membrane bio-reactor (MBR), Nano-filtration (NF), Reverse osmosis (RO) , Waste Water Treatment (WWT).

I. Introduction

WATER uses are manifolds ranging from domestic, agriculture and industries. Among these, chemical industry uses water as a coolant and to generate steam from boilers for use in different industrial processes for the production of different types of products. Depending on its source; water may contain appreciable amount of dissolved calcium and magnesium salts which are a source of hardness in waters. In boilers, water evaporates continuously and the dissolved salts precipitate after reaching saturation stage at equilibrium thus forming a hard scale that deposits on the inner walls of the boiler. There are several disadvantages of these deposits. Among these, the most important is the corrosion which decreases the efficiency of the boiler unit through clogging of pipes, valves and condensers of the unit, decreases the heat transfer rate, causes excessive use of fuel and a danger of explosion. Therefore, wastewater treatment is important for safe operation of boilers. Among the traditional methods of water softening, addition of lime-soda is one process. But this process has some disadvantages i.e leaves more sodium chloride as residue in the raw water and the need of reaction tanks equipped with mechanical stirrers in addition of course to the main chemicals used by the process and coagulant to facilitate filtration of the formed precipitates.

Presently, among the various techniques for softening high hardness waters, membrane separation is a new approach. The membrane separation processes have the unique advantage of not requiring energy to affect phase changes compared to distillation or crystallization. Hence it is economically attractive alternative in view of the fact that it is carried out at reduced energy costs

The objective of this work is to investigate the possibility of using reverse osmosis and nanofiltration processes to improve water quality by removing the major cations and anions such as calcium, magnesium and chloride from wastewater. Also to determine the decline in the performance of RO and NF processes due to membrane fouling.

Review of Literature

RO Technology Application: Membrane technology is playing an increasingly important role in the reclamation of municipal wastewater. Due to the increasing demand for good quality water in urban areas, purification of wastewater has become one of the preferred means of augmenting the water resources [1]. In particular, high quality reclaimed wastewater can be used for industrial customers. For example, it is being used for making boiler feed-water and semiconductor process water. Reverse osmosis (RO) membranes have been proven to successfully treat such water and provide water which exceeds reuse quality requirements. Numerous large-scale commercial membrane plants are now being used to reclaim municipal wastewater. These plants include the 50000 m^3/day West Basin, CA, Kranji 40000 m^3/day in Singapore, and the 32000 m^3/day Bedok plant in Singapore [2]. Additionally, even larger plants are planned such as the 270000 m^3/day plant in Orange County, California and the 380000 m^3/day plant for Sulayabia, Kuwait [3]. The magnitude of these RO-based reclamation plants demonstrates the acceptance that this technology has gained recently.

This work was supported by the King Abdulaziz City for Science and Technology.
Authors are with the National Center for Membrane Technology (NCMT), King Abdulaziz City for Science and Technology (KACST), P. O. Box 6086 Riyadh, Kingdom of Saudi Arabia.
(*e-mail: saljlil@kacst.edu.sa).

A typical process for municipal wastewater consists of primary, secondary and tertiary treatments. The resulting effluent is low in turbidity and can be disinfected for discharge. However, the level of dissolved solids is not reduced by this process and the water is not generally suitable for reuse. When tertiary effluent from a conventional treatment process is supplied to a RO system, it is common to have all forms of fouling - colloidal, biological, scaling and organic fouling. The coatings of foulant will impede water transport through the membranes. Early attempts to treat this water with RO membranes resulted in rapid fouling and required membrane cleaning as frequent as twice per week. This shortened membrane life and greatly increased operating cost.

Since the development of the first practical cellulose acetate membranes in the early 1960's and the subsequent development of thin-film, composite membranes, the uses of reverse osmosis have expanded to include not only the traditional desalination process but also a wide variety of wastewater treatment applications. Several advantages of the RO process that make it particularly attractive for dilute aqueous wastewater treatment include: (1) RO systems are simple to design and operate, have low maintenance requirements, and are modular in nature, making expansion of the systems easy; (2) both inorganic and organic pollutants can be removed simultaneously by RO membrane processes; (3) RO systems allow recovery/recycle of waste process streams with no effect on the material being recovered; (4) RO membrane systems often require less energy and offer lower capital and operating costs than many conventional treatment systems; and (5) RO processes can considerably reduce the volume of waste streams so that these can be treated more efficiently and cost effectively by other processes such as incineration [4, 5, 6, 7, 8, 9]. In addition, RO systems can replace or be used in conjunction with others treatment processes such as oxidation, adsorption, stripping, or biological treatment (as well as many others) to produce a high quality product water that can be reused or discharged. Applications that have been reported for RO processes include the treatment of organic containing wastewater, wastewater from electroplating and metal finishing, pulp and paper, mining and petrochemical, textile, and food processing industries, radioactive wastewater, municipal wastewater, and contaminated groundwater [4, 8, 9, 10].

Contaminated Drinking Water

The ability of RO membranes to remove both inorganic and organic compounds have made these attractive for the treatment of contaminated drinking water supplies [11]. Reverse osmosis processes can simultaneously remove hardness, color, many kinds of bacteria and viruses, and organic contaminants such as agricultural chemicals and trihalomethane precursors. Eisenberg and Middlebrooks [12] reviewed RO treatment of drinking water sources, and they indicated RO could successfully remove a wide variety of contaminants. Chian et al. and Johnston and Lim [13, 14] studied several agricultural chemicals which can contaminate water supplies and found removals were good; however, these adsorbed on the membranes studied. Regunathan et al. [15] reported good removals of the pesticides endrin and methoxychlor as well as trihalomethanes (THMs) with an RO-adsorption system. Nusbaum and Riedinger, Odegaard and Koottatep, and Bhattacharyya and Williams [16, 17, 18] reported that humic and fulvic materials, which are THM precursors, were highly removed by RO membranes. Clair et al. [19] also found excellent removals (>95%) of dissolved organic carbon from natural waters using FT30 membranes. Sorg et al. [20] showed that a RO system could effectively remove radium from contaminated water. Sorg and Love [21] conducted studies with actual groundwater in which only a few of the pollutants being studied were spiked; several different commercial membranes were studied. Most inorganics were highly (>90%) rejected while organic rejection depended upon the organic and membrane studied. Baier et al. [22] studied removal of several agricultural chemicals from groundwater using several different membranes. Rejections ranged from 0% to >94% for the different compounds and membranes studied; pilot plant experiments indicated water fluxes could be maintained over long terms with periodic cleaning. Fronk [23] investigated RO removal of over twenty VOCs and pesticides using several different RO membranes. Average organic removals were 80%. The study indicated that RO could be used to effectively remove both inorganics and organics from drinking water supplies. Taylor et al. [24] found that RO membranes could be used to remove 96% of DOC, 97% of color, 97% of trihalomethane formation potential (THMFP), and 96% of total hardness. Tan and Sudak [25] examined several RO membranes and found all were capable of acceptably removing color from groundwater even over long operating periods.

Municipal Wastewater

The application of RO membranes to the treatment of municipal wastewater has also had some success. Reverse osmosis can remove dissolved solids which cannot be removed by biological or other conventional municipal treatment processes. In addition, RO membranes can also lower organics, color, and nitrate levels. However, extensive pretreatment and periodic cleaning are usually needed to maintain acceptable membrane water fluxes. Early studies [26, 27, 28] showed that high removals of TDS and moderate removals of organics could be achieved. Tsuge and Mori [29] showed that tubular membranes (with a substantial pretreatment system) could remove both inorganics and organics from municipal secondary effluent and produce water meeting drinking water standards. Stenstrom et al. [30] studied municipal wastewater treatment over a 3 year period using tubular cellulose acetate membranes. TDS rejections were 81%,

and TOC rejections were >94%, making the permeate suitable for reuse. However, feed pretreatment was necessary to maintain high water flux levels. Richardson and Argo [31], Allen and Elser [32], Argo and Montes [33], Nusbaum and Argo [34], and Reinhard et al.[35] have discussed municipal wastewater treatment at a large scale plant (Water Factory 21, Orange County, California). The feed to the plant consisted of secondary effluent, and the process was composed of a variety of treatment systems, including RO membranes (several different types) with a 5 MGD capacity. The process reduced TDS and organics to levels that allowed the effluent to be injected into groundwater aquifers used for water supplies. Suzuki and Minami [36] reported studies on use of several RO membranes to treat secondary effluent containing various salts and dissolved organic materials. TDS rejections of up to 99% and TOC rejections as high as 90% were found possible, and fecal coliform group rejections were >99.9%. Losses in water flux over time were noted but could be partially restored by periodic cleaning.

Cséfalvay et al. [37] stated that membrane separations are finding greater use in wastewater treatment because of their efficiency. In order to prove the effectiveness of membrane filtration an applicability study is carried out. Nanofiltration and reverse osmosis membranes are tested under quite different conditions to reduce the chemical oxygen demands (COD) of wastewaters to meet the Council Directive 76/464/EEC release limit. Two kinds of real wastewaters were selected for the investigation. The wastewaters represent extreme different circumstances since the difference between their COD is two orders of magnitude. All of the membranes tested can be applied either to the treatment of wastewater of high COD (pharmaceutical wastewater) or wastewater of low COD (dumpsite leachate), since the different conditions do not change the membrane characteristics. The experimental data show that none of the membranes can decrease the COD to the release limit in one step. However, if two-stage filtrations (nanofiltration followed by reverse osmosis) are accomplished for both of the wastewaters, a total COD reduction of 94% can be achieved. With the application of the two-stage filtration the COD of the wastewater of low COD can be decreased below the release limit but in case of wastewater of the high COD further treatment will be required.

Nanofiltration Applications

Nanofiltration membranes, which have high water fluxes at low pressures, are a recent development that have made possible new applications in wastewater treatment. Nanofiltration membranes are often charged (usually negatively-charged), and, as a result, ion repulsion is the major factor in determining salt rejection. For example, more highly charged ions such as SO_4^{2-} are rejected by most nanofiltration membranes to a greater extent than monovalent ions such as Cl^-. These membranes also reject organic compounds with molecular weights above 200 to 500. These properties have made possible some interesting new applications in wastewater treatment, such as selective separation and recovery of pollutants that have charge differences, separation of hazardous organics from monovalent salt solutions, and membrane softening to reduce hardness and trihalomethane precursors in drinking water sources [8, 38, 39].

Contaminated Drinking Water Supplies

Nanofiltration membranes, although a relatively recent development, have attracted a great deal of attention for use in water softening and removal of various contaminants from drinking water sources. Nanofiltration (NF) processes can reduce or remove TDS, hardness, color, agricultural chemicals, and high molecular weight humic and fulvic materials (which can form trihalomethanes when chlorinated). In addition, NF membranes typically have much higher water fluxes at low pressures when compared with traditional RO membranes used for this application. Conlon [40] reported that FilmTec NF50 membranes could effectively remove color (96%) and TOC (84%), reduce hardness and TDS, and lower trihalomethane formation potential (THMFP) to below regulatory levels. Eriksson [38] and Cadotte et al. [39] also indicated that NF membranes (such as FilmTec NF40, NF50, and NF70) could be used to reduce TDS, hardness, color, and organics. Dykes and Conlon [41], Conlon and McClellan [42], Watson and Hornburg [43], and Conlon et al. [44] have also identified NF as an emerging technology for compliance with THM regulations and for control of TDS, TOC, color, and THM precursors. Clifford et al. [45] discussed the use of NF70 membranes for contaminated groundwater treatment. Removals included 91% for radium-226 and 87% for TDS. Taylor et al. [46] reported that NF70 membranes could allow control of THM formation, DOC, and TDS and produce a high quality product water from an organic contaminated groundwater; they indicated costs of a NF process would be competitive with conventional treatment processes which do not control THMFP. Lange et al. [47] also suggested that NF treatment would be a reliable method of meeting existing and future THM limits compared to chemical treatment alternatives. Amy et al. [48] used NF70 membranes to remove dissolved organic matter from both groundwater (recharged from secondary effluent) and surface water in order to reduce THM precursors; they found that the process was effective in reducing the organics as well as conductivity in both water sources. Tan and Amy [49] showed that NF membranes could remove >88% of color, 51% of TOC, 46% of TDS, and 79% of THMFP from a contaminated water supply. Duranceau et al. [50] and Taylor et al. [51] have reported on the use of NF70 membrane separation of several agricultural chemicals spiked in groundwater. Ethylene dibromide and dibromochloropropane removals averaged 0% and 32%, respectively, while the remaining organics (chlordane,

heptachlor, methoxychlor, and alachlor) were 100% removed. Rejections of TDS were 85% and THMFP were 95%. However, it was also indicated that some of the organics adsorbed on the membrane.

Wastewater Nanofiltration has also been used to remove both organics and inorganics in various wastewaters. Bindoff et al. [52] reported the use of NF membranes to remove color-causing compounds from effluent containing lignins and high salt concentrations in a wood pulping process. Color removals were >98% at water recoveries up to 95% while the inorganics were poorly rejected, allowing the use of low operating pressures (since $\Delta\pi$ was small). Ikeda et al. [53] indicated NF could give high separations of color-causing compounds such as lignin sulphonates in paper pulping wastewaters. Afonso et al. [54] found NF removal (>95%) of chlorinated organic compounds from alkaline pulp and paper bleaching effluents with high water fluxes. Simpson et al. [55] reported the use of NF membranes to remove hardness and organics in textile mill effluents. Gaeta and Fedele [56] also indicated high water recoveries (up to 90%) from textile dye house effluent could be achieved with NF membranes. Perry and Linder [57] discussed the recovery of low molecular weight dyes from high salt concentration effluent. Ikeda et al. [53] and Cadotte et al. [39] reported the use of NF membranes in the treatment of food processing wastewaters. Some specific uses included the desalting of whey and the reduction of high BOD and nitrate levels in potato processing waters (Anonymous, 1988b). Bhattacharyya et al. [58] used NF40 membranes to selectively separate mixtures of cadmium and nickel. Williams et al. [59] and Bhattacharyya and Williams [18] examined NF40 membranes with and without pretreatment by feed preozonation to study removal of various chlorophenols and chloroethanes. TOC rejections up to 90% were possible with ozonation pretreatment. Rautenbach and Gröschl [60] also discussed the separation results of several organics (ranging from methanol to ethylene glycol) by various NF membranes. Chu et al. [61] detailed the use of NF in a process for treating uranium wastewater; NF40 uranium rejections were 97% to 99.9%. Dyke and Bartels [62] discussed the use of NF membranes to replace activated carbon filters for the removal of organics from off shore produced water containing residual oils. The produced waters contained ~1000 mg/L soluble organics (mostly carboxylic acids) and high inorganic concentrations (~15,000 mg/L Na^+ and ~25,000 mg/L Cl^- as well as other dissolved ions). Organic rejections were suitable to meet discharge standards while inorganic rejections were low (<20%), allowing operation at low pressures.

Materials and Methods The experiment was carried at Wastewater Treatment Plant (WTP), National Center for Water technology (NCWT), King Abdulaziz City for Science and Technology (KACST) during 2012-2013.

Analysis of Wastewater Samples The water samples were analyzed for pH, cations and anions. Cations and anions such as chloride, sulphate were determined by using Dionex 300 Ion chromatography. The requirements for this analysis are Dionex ion chromatography with column As-14 (4mm), guard column AS-12, suppressor-ASR-1, fluent mixture of carbonate and bicarbonate, deionized water and nitrogen gas. The results of different Parameters like Cl^-, SO_4^{2-} were obtained in mg/L. The total dissolved solids (TDS) were estimated by using Oven Heraeus Instruments. The pH was measured by using Hach HQ D40.

Experimental set-up: The AWWTU at KACST consists of two units representing two different water treatment technologies such as Reverse Osmosis Unit (RO-Unit) and Nanofiltration (NF).

1. RO-Unit The Pre-treated water from biological unit is desalinized by applying RO-technology. Its water production capacity is 0.12 m^3/h (Fig 1).

Fig 1. Systemic layout of reverse osmosis (RO) unit

2. NF- Unit: The Pre-treated water from biological unit is desalinized by applying NF-technology. Its water production capacity is 0.12 m3/h (Fig 2).

Fig 2. Systemic layout of nanofiltration unit

Results and Discussion The results of RO and NF wastewater treatment technologies containing cations and anions are presented in Fig 3. Only modest rejection was observed for monovalent species as expected with NF. However, rejection/removal of polyvalent cations and anions was strongly observed. On the other hand, strong rejection was observed for monovalent cations and anions by RO [63, 64]. Where the rejection of the species ions (%) $REJ_{(\%)}$ can be calculated as follows:

$$REJ_{(\%)} = \frac{[C_F - C_P]}{[C_F]} * 100$$

where C_F = Concentration of species in the feed and

C_P = Concentration of species in the permate

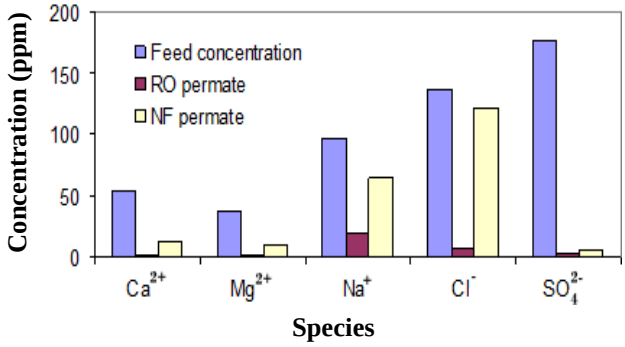


Fig 3. Comparison of the performance of new RO and new NF membranes

Data in Fig 4 shows the results of advanced treatment RO and NF of waste water containing cations and anions when the experiment was carried out using old RO and NF membranes (worked for 3 years). It was found that the rejection of cations and anions decreased for monovalent species. This behavior is expected with old membranes. Also, modest rejection was observed for polyvalent cations and anions. In another hand, modest rejection was observed for monovalent cations and anions by RO. The reason behind this behavior is that the performance of these membranes were decline due to membrane fouling and biofouling. This reason strongly right, because the waste water has different types of bacteria and pathogens and organic material. The organic material adsorbed on the membrane surface and increase the fouling problem.

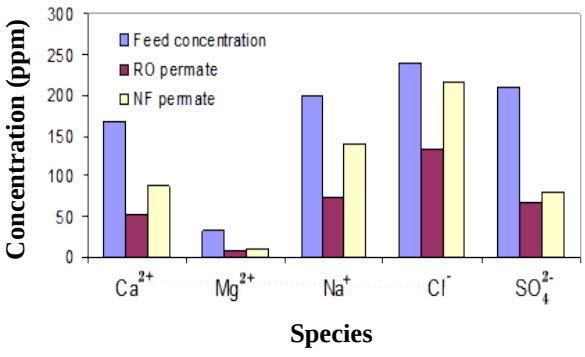


Fig 4. Comparison of the performance of old RO and old NF membranes

The performance of NF and RO for the rejection of TDS is shown in Fig. 5. In general, the TDS rejection decreases with increase in feed concentration due to an increase in concentration polarization at the membrane solution interface [65]. It was observed that TDS rejection was less by NF than RO. This could be attributed to the monovalent ions such as Na^+ which represents the main component in the feed water, therefore, the ability of NF in rejection of the monovalent ions is weak.

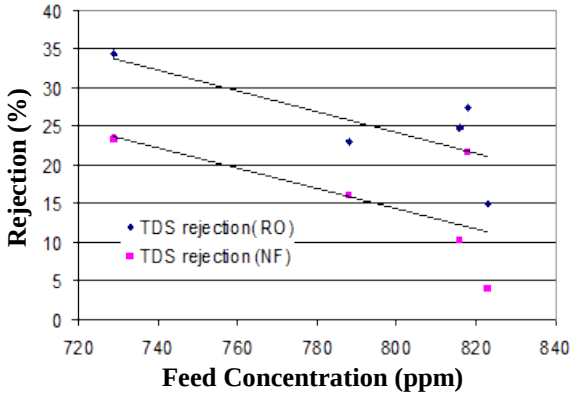
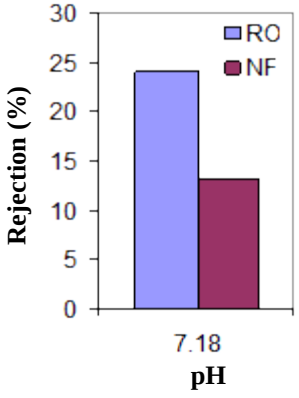


Fig 5. Percentage rejection of TDS versus feed concentration for RO and NF membranes

Fig 6 shows the performance of NF and RO in rejection of TDS at the same pH of feed water. The percentage rejection of the Na^+ ion by RO membrane was higher than by NF membrane. This could be attributed due to the fact that the rejection of monovalent ions such as Na^+ by NF is weak [63, 64]. In addition, the pH did not affect the membrane rejection where the polyamide membrane pH operating ranges between 4-11.



ISBN: 978-1-61804-253-8

Fig 6. Effect of pH on removing sodium ions using RO and NF membranes

Fig 7 shows the results of RO for waste water treatment containing cations and anions by running the experiment using old RO and new RO membranes. The old membrane was 3 years old. It was found that the percentage rejection decreased for all species ions. For example, the percentage rejection of Cl^- ion was 94.4% and after three years the membrane percentage rejection was 43.9. This means that the membrane performance declined due to fouling and biofouling.

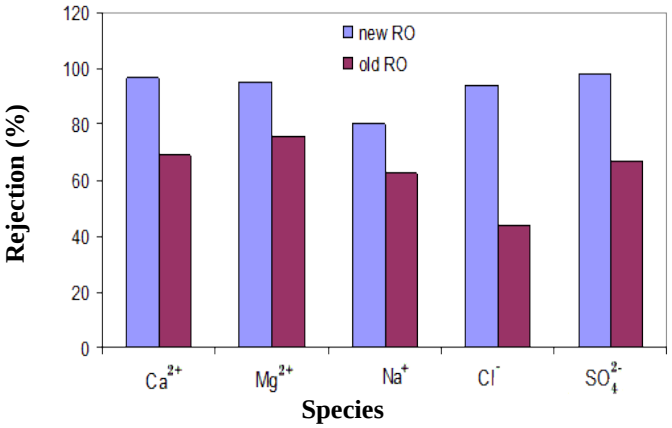


Fig 7. Comparison of the performance of RO membranes during three years operation

Fig 8 shows the results of NF for wastewater treatment containing cations and anions by running the experiment using old NF and new NF membranes. The old membrane worked for 3 years. It was found that, the rejection decreased for polyvalent cations and anions. This behavior was expected with the old membranes. On the other hand, the percent rejection of monovalent cations and anions by NF was not affected. For example, the percentage rejection of Cl^- ion was 11.6% after three years and the membrane rejection was 9.62 %, While the rejection of SO_4^{-2} ions was 97.22 % after three years and the membrane rejection was 61.43 %. A strong rejection was observed for polyvalent cations and anions by NF [64].

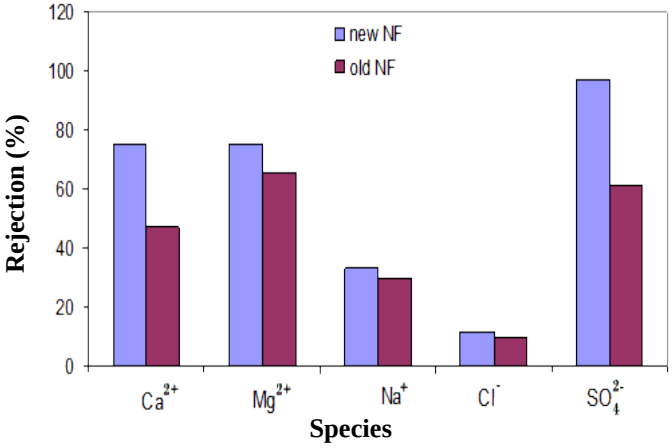


Fig 8. Comparison of the performance of NF membranes during three years operation

Fig 9 and Table 1 show the percent performance decline of RO for treatment of waste water containing cations and anions. It is clear that the decline in the performance of RO was high during three years operation. This decline in performance could be attributed to fouling and biofouling on the surface of the membrane. In conclusion, this promising method ,based on these results, suggest to replace the old membrane by new product. The percentage decline of performance of membrane ($PDPM_{(\%)}$) can be calculate as follows:

$$PDPM_{(\%)} = \frac{NREJ_{(\%)} - OREJ_{(\%)}}{NREJ_{(\%)}} \times 100$$

Where $NREJ_{(\%)}$ = Rejection of new membrane (%) and $OREJ_{(\%)}$ = rejection of old membrane(%)

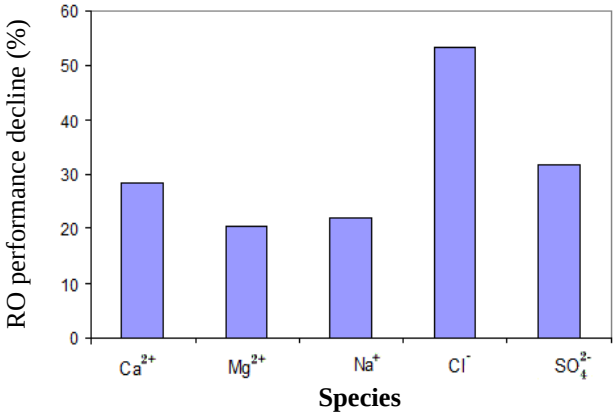


Fig 9. Percentage decline of the performance of RO during three years operation

Fig 10 and Table 2 show the percent decline in the performance of NF for treatment of waste water containing cations and anions. It is clear that the decline in the performance of NF is high during three years operation which might be due to fouling and biofouling on the surface of the membrane. The results of this investigation suggest that the old membrane should be replaced with the new product.

Table 1 : Percentage decline of the performance of RO

Species	$PDPM_{(\%)}$ of RO
Ca^{2+}	28.38
Mg^{2+}	20.39
Na^+	21.98
Cl^-	53.45
SO_4^{2-}	31.73

ISBN: 978-1-61804-253-8

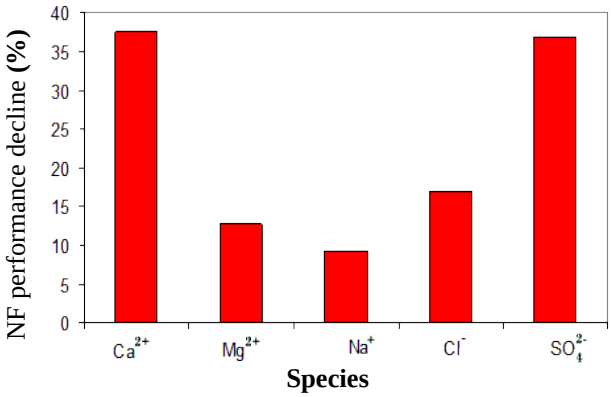


Fig 10. Percentage decline of the performance of NF during three years operation

Table 2 : Percentage decline of the performance of NF

Species	$PDPM_{(\%)}$ of NF
Ca^{2+}	37.67
Mg^{2+}	12.72
Na^{+}	9.37
Cl^{-}	16.97
SO_4^{2-}	36.81

Affecting Factors the decline performance of the membrane such as membrane fouling and biofouling, membrane degradation by oxidation and hydrolysis, mechanical damage, inorganic colloids, adsorbed organics, coagulants, silica scale and other inorganic scale and fouling with waste water. In general, the biofouling and fouling are the major constraints that cause decline in the performance in membranes. In order to find the possible reasons causing decline in membrane performance, membrane autopsy was conducted to collect sheet membrane samples for examination by using energy dispersive x-ray to analyze the fouling deposit. Also, fourier transform infrared spectroscopy was used to identify the components of the deposition that deposit on the membrane. In addition, SEM can be used with the membrane samples to show the advanced fouling on the membrane surface

Control Factors can be used to control fouling on the membrane. These factors are:

1. Efficient pretreatment for the water.
2. Prompt action should be taken in membrane cleaning in the early stages of fouling.
3. Control of bacterial growth by depriving bacteria from nutrition by controlling the organic content in the feed water.
4. Efficient control of membrane fouling by proper sanitization of membrane system by using chlorination or UV.

Conclusions: The nanofiltration (NF) is very effective in removing polyvalent cations and anions such as SO_4^{2-} (where the percent rejection of SO_4^{2-} ions was 97.22 %). While the RO membrane is very effective in removing all ions especially monovalent cations and anions such as Cl^-, where the percent rejection of Cl^- ions was 94.4%. The decline in the performance of RO and NF is significantly high during the first 3-years operation. The main reason behind this is the fouling and biofouling on the surface of the membrane. The suggestion based on these results is to replace the membrane by new one.

Recommendations and suggestions:

1. Change the membrane by new one
2. Observe the performance after fixing a new membrane and make backwash at 10% of decline of the performance.
3. Conduct training for the technicians in RO and NF systems maintenance including the backwash and fouling problem.
4. Cleaning of membrane for recovery of membrane performance.
5. Make a membrane autopsy for the old one to know the exact reason behind the decline of the performance.
6. Use another type of membrane such as ceramic membrane.

REFERENCES

[1] Al-Rehaili, A. M. 1997. Municipal Wastewater Treatment and Reuse in Saudi Arabia. The Arabian Journal for Science and Engineering.22:143-152.

[2] McCray, S., Wytcherley, R., Newbold, D., and Ray, R., 1990. A Review of Wastewater Treatment Using Membranes. Paper Presented at the 1990 International Congress on Membranes and membrane Processes, August 20-24, 1990, Chicago, Illinois.

[3] Al-Shammiri, M. A. Al-Saffar, S. Bohamad and M. Ahmed. 2005. Waste Water quality and reuse in irrigation in Kuwait using microfiltration technology in treatment. Desalination. 185:213-225.

[4] Cartwright, P.S., 1985. Membranes Separations Technology for Industrial Effluent Treatment – A Review. Desalination, 56, 17.

[5] Cartwright, P.S.,1990. Membranes for Industrial Wastewater Treatment - a Technical/Application Perspective. Paper Presented at the 1990

International Congress on Membranes and membrane Processes, August 20-24, Chicago, Illinois.

[6] Cartwright, P.S., 1991. Zero Discharge/Water Reuse - The Opportunities for Membrane Technologies in Pollution Control. Desalination, 83, 225.

[7] Sinisgalli, P., and McNutt, J., 1986. Industrial Use of Reverse Osmosis. Journal AWWA, 47.

[8] Williams, M., Bhattacharyya, D., Ray, R., and McCray, S.,1992. Selected Applications. in Membrane Handbook, W.S.W. Ho and K.K. Sirkar, ed., pp. 312-354, Van Nostrand Reinhold, New York.

[9] Slater, C., Ahlert, R., and Uchrin, C., 1983 a. Applications of Reverse Osmosis to Complex Industrial Wastewater Treatment. Desalination, 48, 171.

[10] Ghabris, A., Abdel-Jawad, M., and Aly, G., 1989. Municipal Wastewater Renovation by Reverse Osmosis, State of the Art. Desalination, 75, 213.

[11] AWWA Membrane Technology Research Committee, 1992. Committee Report: Membrane Processes in Potable Water Treatment. Journal AWWA, 59.

[12] Eisenberg, T., and Middlebrooks, E.,1986. Reverse Osmosis Treatment of Drinking Water. Butterworth, Boston.

[13] Chian, E., Bruce, W., and Fang, H., 1975. Removal of Pesticides by Reverse Osmosis. Environmental Science and Technology, 9, 364.

[14] Johnston, H., and Lim, H., 1978. Removal of Persistent Contaminants from Municipal Effluents by Reverse Osmosis. Report 85, Ontario Mining Environmental, Ontario.

[15] Regunathan, P., Beauman, W., and Kreusch, E., 1983. Efficiency of Point-of-Use Devices. Journal AWWA, 75, 42.

[16] Nusbaum, I., and Riedinger, A., 1980. Water Quality Improvement by Reverse Osmosis. in Water Treatment Plant Design, R. Sanks, ed., Ann Arbor Science, Ann Arbor, Michigan.

[17] Odegaard, H., and Koottatep, S., 1982. Removal of Humic Substances from Natural Waters by Reverse Osmosis. Water Research, 16, 613.

[18] Bhattacharyya, D., and Williams, M., 1992 a. Separation of Hazardous Organics by Low Pressure Reverse Osmosis Membranes - Phase II, Final Report. EPA Report, EPA/600/2-91/045.

[19] Clair, T., Kramer, J., Sydor, M., and Eaton, D., 1991. Concentration of Aquatic Dissolved Organic Matter by Reverse Osmosis. *Water Research, 25,* 1033.

[20] Sorg, T., Forbes, R., and Chambers, D., 1980. Removal of Radium 226 from Sarasota County, FL, Drinking Water by Reverse Osmosis. Journal AWWA, 72, 230.

[21] Sorg, T., and Love, Jr., O., 1984. Reverse Osmosis Treatment to Control Inorganic and Volatile Organic Contamination. EPA Report, EPA 600/D-84-198.

[22] Baier, J., Lykins, Jr., B., Fronk, C., and Kramer, S.,1987. Using Reverse Osmosis to Remove Agriculture Chemicals from Groundwater. Journal AWWA, 55.

[23] Fronk, C., 1987. Removal of Low Molecular Weight Organic Contaminants from Drinking Water Using Reverse Osmosis Membranes. EPA Report, EPA/600/D-87/254.

[24] Taylor, J., Thompson, D., and Carswell, J., 1987. Applying Membrane Processes to Groundwater Sources for Trihalomethane Precursor Control. Journal AWWA, 72.

[25] Tan, L., and Sudak, R., 1992. Removing Color from a Groundwater Source. Journal AWWA, 79.

[26] Cruver, J.,1976. Waste-treatment Applications of Reverse Osmosis. Transactions ASME, 246.

[27] Fang, H., and Chian, E., 1976. Reverse Osmosis Separation of Polar Organic Compounds in Aqueous Solution. Environmental Science and Technology, 10, 364.

[28] Lim, M., and Johnston, H., 1976. Reverse Osmosis as an Advanced Treatment Process. Journal WPCF,48, 1820.

[29] Tsuge, H., and Mori, K.,1977. Reclamation of Municipal Sewage by Reverse Osmosis. Desalination ,23, 123.

[30] Stenstrom, M., Davis, J., Lopez, J., and McCutchan, J., 1982. Municipal Wastewater Reclamation by Reverse Osmosis - A 3-year Case Study. Journal WPCF, 54, 43.

[31] Richardson, N., and Argo, D., 1977. Orange County's 5 MGD Reverse Osmosis Plant. Desalination, 23,563.

[32] Allen, P., and Elser, G., 1979. They Said it Couldn't be Done - the Orange County, California Experience. Desalination, 30, 23.

[33] Argo, D., and Montes, J., 1979. Wastewater Reclamation by Reverse Osmosis. Journal WPCF, 51, 590.

[34] Nusbaum, I., and Argo, D., 1984. Design, Operation, and Maintenance of a 5-mgd Wastewater Reclamation Reverse Osmosis Plant. in Synthetic Membrane Processes: Fundamentals and WaterApplications, G. Belfort, ed., Academic Press, New York.

[35] Reinhard, M., Goodman, N., McCarty, P., and Argo, D., 1986. Removing Trace Organics by Reverse Osmosis Using Cellulose Acetate and Polyamide Membranes. Journal AWWA, 163.

[36] Suzuki, Y., and Minami, T., 1991. "Technological Development of a Wastewater Reclamation Process for Recreational Reuse: An Approach to

Advanced Wastewater Treatment Featuring Reverse Osmosis Membrane", Water Science and Technology, 23, 1629.

[37] Cséfalvay, E., P. M. Imre and P. Mizsey. 2008. Applicability of nanofiltration and reverse osmosis for the treatment of wastewater of different origin. Central European Journal of Chemistry.6(2):277-283.

[38] Eriksson, P., 1988. Nanofiltration Extends the Range of Membrane Filtration. Environmental Progress,7, 58.

[39] Cadotte, J., Forester, R., Kim, M., Petersen, R., and Stocker, T., 1988. Nanofiltration Membranes Broaden the Use of Membrane Separation Technology. Desalination, 70, 77.

[40] Conlon, W., 1985. Pilot Field Test Data for Prototype Ultra Low Pressure Reverse Osmosis Elements. Desalination, 56, 203.

[41] Dykes, G., and Conlon, W.,1989. Use of Membrane Technology in Florida. Journal AWWA, 43.

[42] Conlon, W., and McClellan, S., 1989. Membrane Softening: A Treatment Process Comes of Age. Journal AWWA, 47.

[43] Watson, B., and Hornburg, C.,1989. Low-Energy Membrane Nanofiltration for Removal of Color, Organics and Hardness from Drinking Water Supplies. Desalination, 72, 11.

[44] Conlon, W., Hornburg, C., Watson, B., and Kiefer, C., 1990. Membrane Softening: The Concept and its Application to Municipal Water Supply. Desalination, 78, 157.

[45] Clifford, D., Vijjeswarapu, W., and Subramonian, S., 1988. Evaluating Various Adsorbents and Membranes for Removing Radium from Groundwater. Journal AWWA, 94.

[46] Taylor, J., Mulford, L., and Duranceau, S., 1989 a. Cost and Performance of a Membrane Pilot Plant. Journal AWWA, 52.

[47] Lange, P., Laverty, P., Edwards, E., and Watson, I., 1989. THM Precursor Removal and Softening - FT. Myers 12 MGD RO Membrane Plant, Florida USA. Desalination, 76, 39.

[48] Amy, G., Alleman, B., and Cluff, C., 1990. Removal of Dissolved Organic Matter by Nanofiltration. Journal of Environmental Engineering, 116, 200.

[49] Tan, L., and Amy, G.L., 1991. Comparing Ozonation and Membrane Separation for Color Removal and Disinfection By-Product Control. Journal AWWA, 74.

[50] Duranceau, S., Taylor, J., and Mulford, L., 1992. SOC Removal in a Membrane Softening Process. Journal AWWA, 68.

[51] Taylor, J., Durancean, S., Mulford, L., Smith, D., and Barrett, W., 1989 b. SOC Rejection by Nanofiltration. EPA Report, EPA/600/2-89/023.

[52] Bindoff, A., Davies, C., Kerr, C., and Buckley, C., 1987. The Nanofiltration and Reuse of Effluent from the Caustic Extraction Stage of Wood Pulping. Desalination, 67, 453.

[53] Ikeda, K., Nakano, T., Ito, H., Kubota, T., and Yamamoto, S., 1988. New Composite Charged Reverse Osmosis Membrane. Desalination, 68, 109.

[54] Afonso, M., Geraldes, V., Rosa, M., and De Pinho, M., 1992. Nanofiltration Removal of Chlorinated Organic Compounds from Alkaline Bleaching Effluents in a Pulp and Paper Plant. Water Research,26, 1639.

[55] Simpson, M., Kerr, C., and Buckley, C., 1987. The Effect of pH on the Nanofiltration of the Carbonate System in Solution. Desalination, 64, 305.

[56] Gaeta, S., and Fedele, U., 1991. Recovery of Water and Auxiliary Chemicals from Effluents of Textile Dye Houses. Institution of Chemical Engineers Symposium Series, 3, 183.

[57] Perry, M., and Linder, C., 1989. Intermediate Reverse Osmosis Ultrafiltration (RO UF) Membranes for Concentration and Desalting of Low Molecular Weight Organic Solutes. Desalination, 71, 233.

[58] Bhattacharyya, D., Adams, R., and Williams, M., 1989. Separation of Selected Organics and Inorganic Solutes by Low Pressure Reverse Osmosis Membranes. in Biological and Synthetic Membranes, D. Butterfield, ed., Alan R. Liss, New York.

[59] Williams, M., Deshmukh, R., and Bhattacharyya, D., 1990. Separation of Hazardous Organics by Reverse Osmosis Membranes. Environmental Progress, 9, 118.

[60] Rautenbach, R., and Gröschl, A., 1990 b. Separation Potential of Nanofiltration Membranes. Desalination,77, 73.

[61] Chu, M., Tung, C., and Shieh, M., 1990. A Study on Triple-Membrane-Separator (TMS) Process to Treat Aqueous Effluents Containing Uranium. Separation Science and Technology, 25, 1339.

[62] Dyke, C., and Bartels, C., 1990. Removal of Organics from Offshore Produced Waters Using Nanofiltration Membrane Technology. Environmental Progress, 9, 183.

[63] Khedr, M. G.," Nanofiltration and low energy reverse osmosis for rejection of radioactive isotopes and heavy metal cations from drinking water sources", Desalination and Water Treatment, 342, 2009.

[64] Albino, K., Donald, B.," Brackish and seawater desalting, reverse osmosis technology, B. S.

Parekh, ed., Marcel Dekker, New York, 1988, 271, 2000.

[65] Khedr, M. G.," Membrane fouling problems in reverse osmosis desalination applications "International Desalination & Water Reuse, 10 (3), 8, 2000.

Study of seepage for small homogeneous earth dams

Marius Lucian Botos

***Abstract*—** Constructed in order to regulate the water level in one section of a river, dams, regardless of their type, are subjected to the permanent or non-permanent action of water. Earth fill dams' cross section geometry is designed based on the study of stability for the upstream and downstream slope. The design is influenced by several factors; type of soil used; drainage solution; and in particular the seepage. The calculation of seepage is determining the delimitation of the saturated soil from the unsaturated one. In the particular case of small dams the water level in the reservoir varies quickly in time, hence seepage is influenced by this variation, as well as by the type of soil used, and the presence or lack of a drainage system.

Current paper presents the results of research carried out to determine seepage characteristics for different exploitation strategies of such reservoirs, evolution of saturated area if the maximum water level is maintained constant for a long period of time. Using results from such a study shows that seepage has an important role when evaluating the slope stability or when readings at piezometers are used to evaluate behavior of existing dams**.**

***Keywords*—** seepage, slope stability, small dam.

I. Introduction

In Romania there is currently great potential to increase the volume of fresh water stored in reservoirs without the need for new dams construction. Issues related to environmental protection and private property in recent years cause this type of construction to be increasingly more difficult to start with. An important number of non-permanent storage dams are currently with a single usage: flood wave attenuation. Adding new use by permanent accumulation poses a number of problems related to the behavior of the dam to the new conditions.

Non permanent storage dams fall into the category of small dams [12], in Romania 92% of them are earth fill homogeneous dams with heights up to 10 meters [2].

To understand in detail the behavior of this type of works, this paper presents a comprehensive study on the changes, which occur from seepage point of view, for most cases possible in practice. Understanding behavior on seepage opens the way to solve problems related to the stability of slopes. Most of studies on unsteady seepage through earth dams focus on large dams with cores made of clay, subject to slow variations of the water levels up and downstream.

Small dams are characterized by small heights and/or small storage volumes. Variation of water level in the lake is difficult to control and generally flood is attenuated without performing maneuvers or pre emptying. Thought to take and mitigate floods in the superior basins of rivers, characterized by hydrographs with short growth times, causing rapid variations flow and levels. This type of load automatically imposes, that the seepage has to be modeled considering the unsteady unsaturated case.

II. Governing equations

The present problem to be solved consists in a homogeneous earth isotropic dam founded on a layer of soil with same characteristics as the embankment. The impervious layer is placed at same depth as the dam's height. The pore pressure based formulation of Richards equation for two dimensional flow of water through a homogeneous unsaturated soil is according to (1):

$$\frac{\partial}{\partial x}\left(k_x(\psi)\frac{\partial \psi}{\partial x}\right)+\frac{\partial}{\partial y}\left(k_y(\psi)\frac{\partial \psi}{\partial y}\right)=C(\psi)\frac{\partial \psi}{\partial t} \quad (1)$$

$k_x(\psi), k_y(\psi), k_z(\psi)$ functions of permeability depending on of the pore water pressure (2).

The specific moisture capacity:

$$C(\psi)=\frac{\partial \theta}{\partial \psi} \quad (2)$$

θ - volumetric water content.

For the definition permeability functions are numerous examples in the scientific literature Childs and Colins [6]; Burdine [5]; Mualem and Dagan [16]; Kosugi [16].

For the estimation of permeability function is used the Burdine model :

$$k_r = S_e^l\left[1-\left(1-S_e^{1/m}\right)^m\right] \quad (3)$$

where:

k_r - relative permeability;

S_e - the effective saturation:

$$S_e=\frac{\theta-\theta_r}{\theta_s-\theta_r} \quad (4)$$

θ_s is the saturated moisture content and θ_r is the residual moisture content of the soil.

M. L. Botos. is with the Technical University of Cluj-Napoca, Faculty of Civil Engineering, 400020, Cluj-Napoca (phone: +40766245888; e-mail: marius.botos@mecon. utcluj.ro).

The relationship between pressure head and effective saturation derived from the van Genuchten model [17] :

$$S_e = \frac{1}{\left[1+(-\alpha\psi)^n\right]^m} \quad (5)$$

For van Genuchten model the equation of specific moisture capacity becomes:

$$C = (\theta_s - \theta_r)mn(-\alpha\psi)^{n-1}\left[1+(-\alpha\psi)^n\right]^{-m-1} \quad (6)$$

Appling the method of weighted residuals, forcing the weighted residuals to be equal to zero in each Gauss node and considering weighted function to be the same as interpolating functions a system of ordinary differential equations is obtained [13]:

$$\underset{global}{[C(\psi)]}\begin{Bmatrix}\frac{\partial\psi_1}{\partial t}\\ \vdots \\ \frac{\partial\psi_n}{\partial t}\end{Bmatrix} + \underset{global}{[K(\psi)]}\begin{Bmatrix}\psi_1\\ \vdots\\ \psi_n\end{Bmatrix} = \{0\} \quad (7)$$

$\underset{global}{[K(\psi)]}$ - global conductance matrix; $\underset{global}{[C(\psi)]}$ global capacitance matrix.

System of equations can be solved using backward difference method [13] :

$$([C(\psi)] + \omega\Delta t[K(\psi)])\{\psi\}_{t+\Delta t} = ([C(\psi)] - (1-\omega)\Delta t[K(\psi)])\{\psi\}_t \quad (8)$$

III. Initial conditions

In order to start calculations some initial conditions are specified on the boundary for all types of dams, for all types of exploitation regimes. For the initial conditions defined steady state of seepage was taken in consideration. The Dirichlet boundary condition specifies the pressure head on some part of the boundary, whereas the Neuman condition specifies the flux on other part of the boundary. Neuman conditions were not specified, no nodal fluxes were considered. On the upstream wetted nodes were assigned with pressure heads, downstream toe was considered dry and pressure heads were defined as zero. The initial condition consists in obtaining distribution of the pressure head and the saturation throughout the solution domain at the start of the solution history.

IV. Boundary conditions

In all the underlying assumptions that we have considered the water level downstream is equal to nil, water level in the lake will linearly or suddenly vary. For the dam with 2 m height is considered an increase in the total time of 6 hours, 5 m for the 12 hours and the dam with a height of 10 m from the calculation of the total 24 hours. For each height there were considered two types of operating:

1. The dam is considered with nonpermanent storage and water level variation occurs from 0 to the maximum height of the water in the lake, while the variation period is listed above.

2. The dam with permanent storage, the water level variation is linear, from to a normal retention level considered 1/3 of the height of the dam, up to the maximum.

V. Boundary conditions

We studied in detail the homogeneous dams with 2, 5 and 10 meters height. There have been proposed to study cross-sections with the following characteristics:

Table 1. Slopes of the embankments

Soil	***Upstream slope***	***Downstream slope***
Sandy clay	3:1	2.5:1

The crest width was considered as proposed by Lewis [15] the foundation soil is the same type of material used in the dam's body and its thickness equal to the height of the dam.

Table 2. Parameters of dams with slopes of 1:3 and 1:2.5

Dams height 2 m			***Dams height 5 m***			***Dams height 10 m***		
Crest width B=2.5m			***Crest width B=3.25 m***			***Crest width B=4.20 m***		
No. nodes	No. elem.	T [h]	No. nodes	No. elem.	T [h]	No. nodes	No. elem.	T [h]
1260	2299	6	2256	4233	12	1901	3573	24

To solve the problem of unsteady seepage was necessary to choose a model for the permeability and water retention functions.

In order to solve unsteady unsaturated seepage, for the water retention functions was chosen the model proposed by van Genuchten and for the permeability functions the model proposed by Burdine [5]. Parameters for the soil were obtained using RETC software [18].

Table 3. Parameters characterizing seepage for the unsaturated zone

CL sandy clay
θ_R=0.0672; θ_S=0.3963 α=2.4; n=1.3348; m=0.2508
K_S=1.416x10^{-6} m/s

VI. Results and Discussion

The models were composed in the gmsh.exe program, meshing is uniform throughout the area, the size of the elements as the result of a sensitivity analysis around 0.15 square meters, the elements used to mesh are triangular and linear. In order to verify how the type of the water level change in the lake (linear or suddenly) influences the results, one of the hypotheses considers at time $t >\approx 0,0\,t$ water level suddenly increases to maximum height. To solve the problem of unsteady seepage through unsaturated soils was used a variation of MNPNS.exe program [1]. Systems of nonlinear equations are solved by the substitution method, for the solution to be acceptable; the required tolerance for two successive iterations is 0.5%.

A. Sandy clay dam with height of 2 meters

Considering dam with height of two meters may seem a little unfortunate, but the results obtained may prove wrong. For

different type of dam exploitation rules, with permanent or non permanent storage we obtained different final result for the end of time period considered. Results are presented in Fig. 1 and Fig. 2.

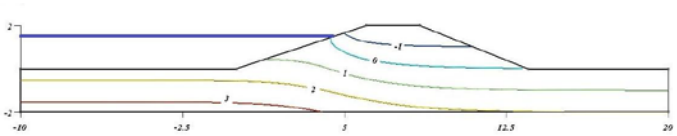


Fig. 1. Linear water level rise from 0 to 1.50 m

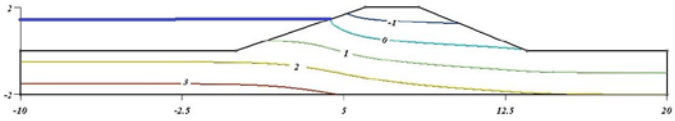


Fig. 2. Linear water level rise from 0.60 to 1.50 m

The evolution of saturated area inside the embankment for different positions is shown in the Fig. 3 and Fig. 4.

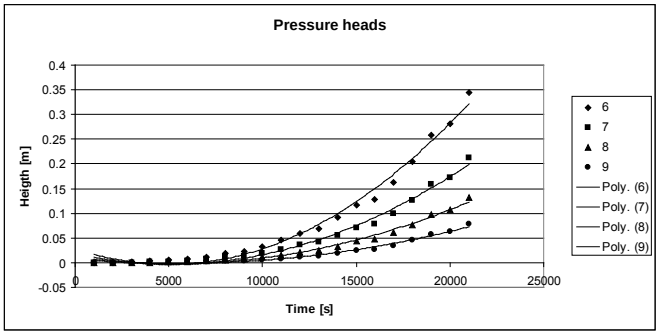


Fig. 3. Evolution of pressure head for different piezometer positions for dams with non permanent storage

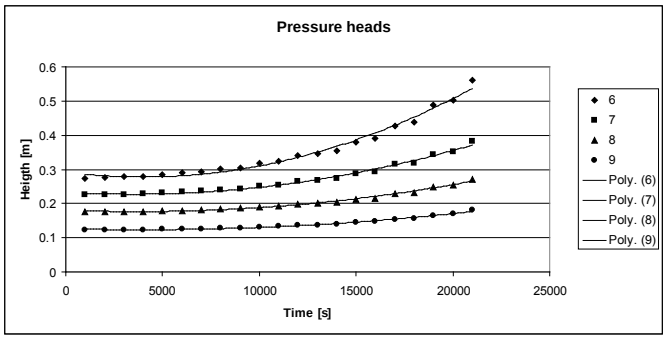


Fig. 4. Evolution of pressure head for different piezometer positions for dams with permanent storage

For a better understanding of the importance of unsteady seepage analysis the results are shown for the same positions of virtual piezometers, compared to the steady seepage analysis results. For the steady seepage analysis the water level in the lake was considered at the maximum. Considering the results in maters of percentage, allows us to understand, that the unsteady seepage analysis is important to be conducted even for small dams.

For the each piezometer position five different results are presented in Fig. 5. First is the water level in the piezometer obtained considering linear raise of the water level for nonpermanent lake, second for a sudden raise of water in the lake for the same type of storage.

The middle column is the benchmark, in our case the result for the steady seepage for the maximum level. Last two columns are representing the results for the permanent lake, first one for the linear variation of the level and last one for the sudden raise of boundary conditions.

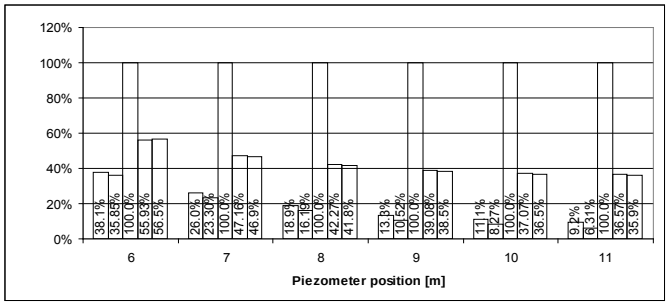


Fig. 5. Comparison of the piezometric surface in different scenario compared with steady seepage

B. Sandy clay dam with height of 5 meters

For the five meters height dam results of pore pressure distribution and the position of the virtual piezometers are presented in Fig. 6 and Fig. 7.

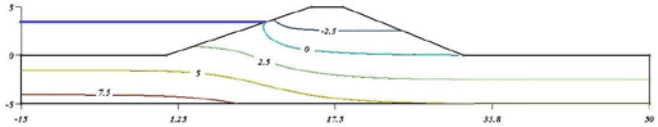


Fig. 6. Linear water level rise from 0 to 4.50 m

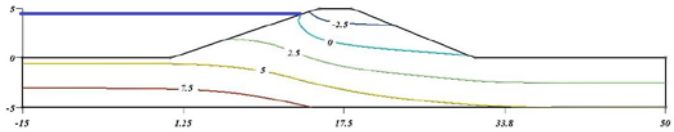


Fig. 7. Linear water level rise from 1.60 to 4.50 m

Evolution of the saturated zone in the embankment shown in Fig. 8 and Fig. 9 for the different exploitation scenarios are exposing the fact that important changes occur in case of dams with nonpermanent storage after first half of the interval, but only in the upstream part of the dam.

ISBN: 978-1-61804-253-8

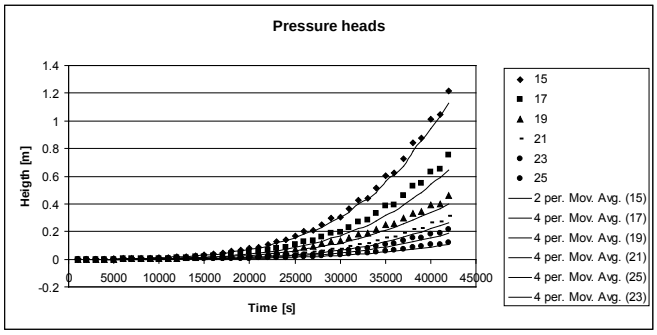


Fig. 8. Evolution of pressure head for different piezometer positions for dams with non permanent storage

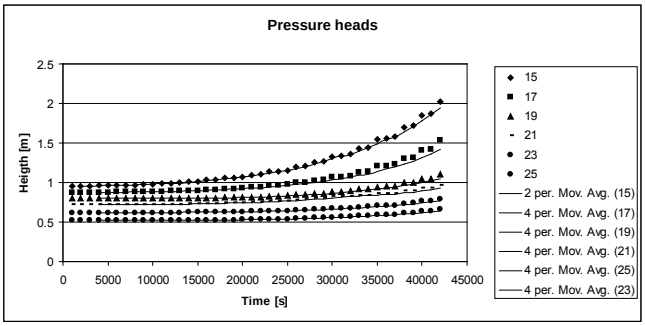


Fig. 9. Evolution of pressure head for different piezometer positions for dams with permanent storage

Like the results obtained in the other cases, differences obtained for different types of loading in the lake (for the chosen time interval) are showing negligible differences for the final results.

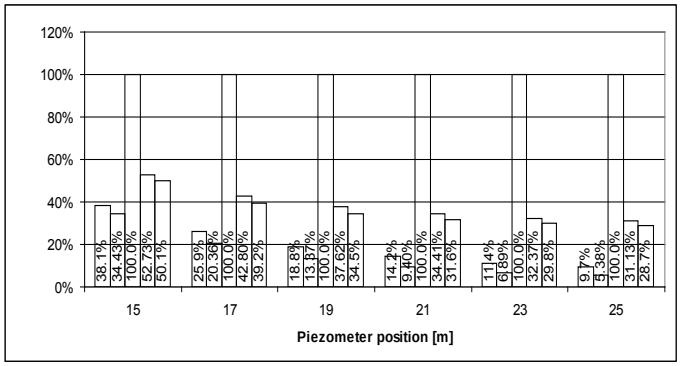


Fig. 10. Comparison of the piezometric surface in different scenario compared with steady seepage

The results obtained by modeling unsteady seepage through unsaturated soils reveal significant differences in medium and small dams (2 to 5 m), where the water level in the drained dams with nonpermanent reservoirs is between 38% and 10% of the level achieved the steady seepage and between 56% and 32% for the dams with permanent reservoir.

C. Sandy clay dam with height of 10 meters

Characterized by a bigger volume to be saturated, dams with 10 meters height in order to be in the category of small dams need to have a volume smaller than 4 millions cubic meters. This means that total growth times for the hydrographs can not be more than 24 hours. The model considers that as the growth time; the final and the evolution of results are presented in Fig. 11 and Fig. 12.

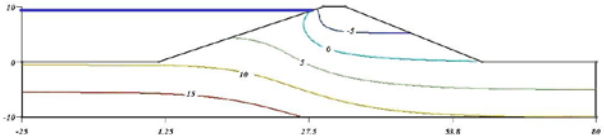


Fig. 11. Linear water level rise from 0 to 9.50 m

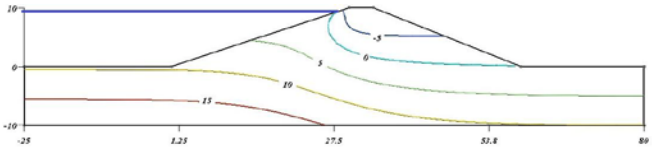


Fig. 12. Linear water level rise from 3.3 to 9.50 m

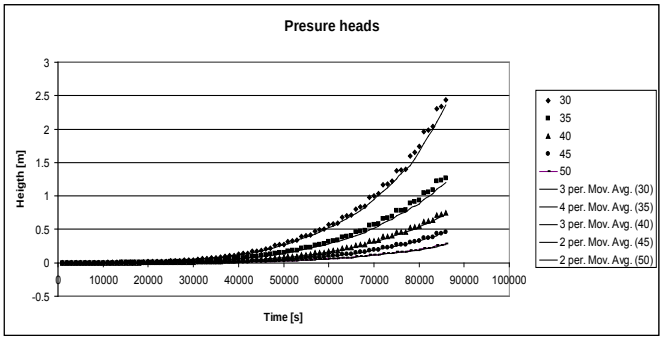


Fig. 13. Evolution of pressure for different piezometer positions for dams with non permanent storage

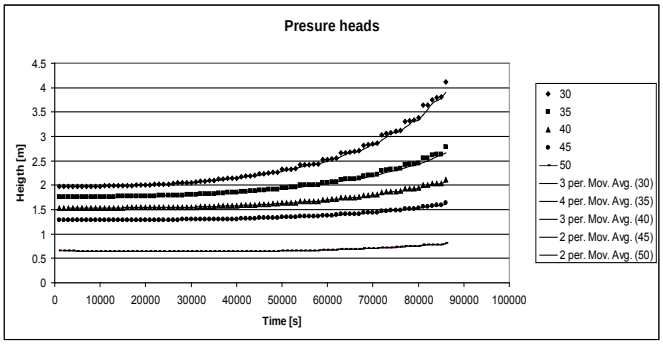


Fig. 14. Evolution of pressure heads for different piezometer positions for dams with permanent storage

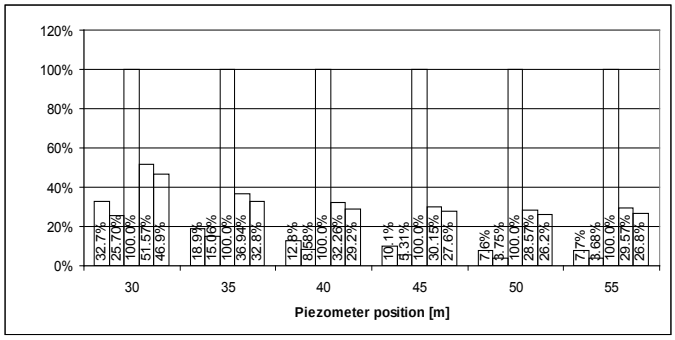


Fig. 15. Comparison of the piezometric surface in different scenario compared with steady seepage

The way the water level variation in lake is considered (linearly or suddenly over the analyzed period) reveals that the differences for dams with the height of 10 meters are 7% for the upstream area of the dam and under 4-5% toward downstream.

Regardless of the height of the dam, the results obtained show a significant difference compared to those obtained by classical methods. The results presented in this paper are only a part of the study conducted on the infiltration for majority types of soils used for dams in homogeneous solution: silty sand, clayey sand, silty clay and fat cay. The main aspect that influences the way the saturated zone evolves is the cross-sectional dimensions. The percentage values reported to the steady state seepage remain close in value for each height, varying within 5%. For dams with non permanent storage: heights of 2 meters between 39% for the upstream piezometers to 10% for the downstream piezometers, 5 meters heights between 37% and 8% and for 10 meters dam between 28% and 6%. For dams with permanent storage: 2 meters height between 59% and 37%, 5 meters 53% and 33%, and for 10 meters values between 45 and 26%.

VII. Conclusion

Comparison of water levels in piezometers in relation to the results obtained if the seepage is considered steady at maximum levels during floods can not provide the real behavior of the dam and gives a false sense of safety. The way water flow is calculated through earth fill dams, influences the stability safety factor of downstream slope of embankment.

If the model is considering the seepage in steady conditions, the readings at specific inspection equipment will indicate significantly different values. Costly investments can be made in order to replace the inspection equipment with no result. Confusion intervenes and excessive caution can be imposed in the exploitation rules, caution that may lead to a decrease in the degree of flood attenuation and unnecessary damages downstream. This prudence can lead to the initiation of expensive investments such waterproofing or oversized draining elements.

References

[1] Botos M.L. Solving steady and unsteady seepage problems using custom programs, *C60 International Conference*, Cluj-Napoca, 2013

[2] Botos M.L. Small dams with non-permanent reservoirs in Romania, *C60 International Conference*, Cluj-Napoca, 2013.

[3] Brooks R.H , Corey A.T. Hydraulic properties of porous media *Hydrology Paper No. 3*, Colorado State University, Fort Collins, Colorado 1964.

[4] Brutsaert W. Some methods of calculating unsaturated permeability *Transactions of the American Society of Agricultural Engineers*, 10, 1966, pp. 400–404

[5] Burdine, N.T. Relative permeability calculations from pore-size distribution data. *Trans. Am. Inst. Min. Metall.* Pet. Eng. 198:71.77. 1953.

[6] Childs, E.C., and N. Collis-George. The permeability of porous materials. *Proc. R. Soc. London*, Ser. A. 201:392.405. 1950.

[7] Fredlund D.G, A. Xing Equations for the soil–water characteristic curve *Canadian Geotechnical Journal*, 31 (4), 1994, pp. 521–532

[8] Fredlund D.G. , A. Xing, S. Huang Predicting the permeability function for unsaturated soil using the soil–water characteristic curve *Canadian Geotechnical Journal*, 31 (3), 1994, pp. 521–532

[9] Fredlund D.G. , D. Sheng, J. Zhao Estimation of soil suction from the soil–water characteristic curve *Canadian Geotechnical Journal*, 48, 2011, pp. 186–198

[10] Gardner W.R. Some steady state solutions of the unsaturated moisture flow equation with application to evaporation from a water table *Soil Science*, 85 (4), 1958, pp. 228–232

[11] Green R.E. , J.C. CoreyCalculation of hydraulic conductivity: a further evaluation of some predictive methods *Soil Science Society of America Proceedings*, 35, 1971, pp. 3–8

[12] ICOLD. SMALL DAMS *Design, Surveillance and Rehabilitation*, 2011

[13] Istok J. D. , *Groundwater Modelling by the Finite Element Method*, Water Resources Monograph Series, Volume 13, 1989.

[14] Kosugi, K. . General model for unsaturated hydraulic conductivity for soils with lognormal poresize distribution. *Soil Sci. Soc. Am. J*., 1999.

[15] Lewis B. – *Farm Dams – Planning, Construction and Maintenance*, National Library of Australia Cataloguing; 2001.

[16] Mualem, Y., and G. Dagan. Hydraulic conductivity of soils: Unified approach to the statistical models. *Soil Sci. Soc. Am.*1978.

[17] van Genuchten M.T A closed-form equation for predicting the hydraulic conductivity of unsaturated soils Soil *Science Society of America Journal*, 44, 1980,pp. 892–898

[18] Yates, S.R., M.Th. van Genuchten, A.W.Warrick, and F.J. Leij. Analysis of measured, predicted, and estimated hydraulic conductivity using the RETC computer program. *Soil Sci. Soc. Am.*1992.

Effect of Fuels on Gas Turbine Can-Type Combustor using CFD Code

Guessab A., Aris A. Benabdallah T. and Chami N.

***Abstract*—**This paper describes reacting flow analysis of a gas turbine combustion system. The method is based on the solution of Navier-Strokes equations using finite volume method. The turbulence effects are modeled through the renormalization group k-ε model. The method has been applied to a practical gas turbine can-type combustor. The numerical models used for natural gas and biogas fuel combustion consist of the eddy dissipation model for non-premixed gas combustion and the Two-step of Westbrook and Dryer was used to compute the rate of fuel oxidation. The combustion system includes swirler vane passages, fuel nozzles cone and all holes in primary zone, dilution zone. The total geometry has been created using the pre-processors GAMBIT and SolidWorks and the meshing has been done using GAMBIT, and the analysis carried out in ANSYS FLUENT solver. These models are use to analysis the effects of alternative fuels on the combustion emission characteristics of air-fuel mixture are examined. The composition of the fuel injected in can combustor was changed from natural gas (100% CH_4) to biogas1 (90% CH_4+10% CO_2), Biogas2 (75% CH_4+25% CO_2), and Biogas 3 (70% CH_4+30% CO_2). The effects of these variations on the structure (temperature, NO and CO_2 mass fractions) of the can-type combustor gas turbine chamber are calculated. The numerical analysis has shown that the biogas fuel allows a reduction of about 35% on the NO emissions and about 20% on the CO_2 emissions.

***Keywords*—** Can-type combustor, Biogas, EDM, Natural gas, NO formation, RNG k-ε model.

I. Introduction

The turbines are widely used in modern industry to deliver shaft power or thrust power. The combustor is part of the gas turbine. The main goal of the combustor design is lower emissions with less volume. Flow field simulation in the combustor is a challenging subject to both academics and industries. It is of commercial importance to understand and to predict various phenomena in the combustor. Gas turbine combustion systems need to be designed and developed to meet many mutually conflicting design requirements, including high combustion efficiency over a wide operating envelope and low NOx emission, low smoke, low lean flame stability limits and good starting characteristics; low combustion system pressure loss, low pattern factor, and sufficient cooling air to maintain low wall temperature levels and gradients commensurate with structural durability. The use of alternative fuels in gas turbine combustion chambers can provide advantages such as lower environmental impact. Combustion phenomenon will be accomplished effectively by the development of low emission combustor. One of the significant factors influencing the entire Combustion process is the mixing between a swirling primary air and the non-swirling fuel. To study this fundamental flow, the chamber had to be designed in such a manner that the combustion process to sustain itself in a continuous manner and the temperature of the products is sufficiently below the maximum working temperature in the turbine. This study is used to develop the effective combustion with low unburned combustion products by adopting the concept of high swirl flow and motility of holes in the secondary chamber. The proper selection of a swirler is needed to reduce emission which can be concluded from the emission of NOx and CO_2. The capture of CO_2 is necessary to mitigate CO_2 emissions from natural gas. Thus the suppression of unburned gases is a meaningful objective for the development of high performance combustor without affecting turbine blade temperature.

The combustion of methane-air mixture in gas turbine can-type combustion chamber was experimented numerically by Pathan *et al.* [1], Koutmos and McGuirk [2], Ghenai [3], Eldrainy *et al.* [4].

In this paper, modifications of the gas turbine can-type combustion chamber are investigated in order to reduce NO_x emissions with biogas operation. The impact of modifications on air-fuel mixing, temperature field and NO volumetric formation rate was evaluated using a CFD 3D RANS reactive procedure validated by comparison with the natural gas data. The FLUENT commercial code was employed for numerical simulations. Simulations show that the proposed modifications could reduce NO emissions up to 30%.

II. Model Description and Simulation Detail

The basic geometry of the gas turbine can-type combustor chamber is shown in Fig. 1. The size of the combustor is 590 mm in the *Z* direction, 250 mm in the *Y* direction, and 230 mm in the *X* direction (Fig. 2). The primary inlet air is guided by vanes to give the air a swirling velocity component. The total surface area of primary main air inlet is 57 cm^2. The fuel is injected through six fuel inlets in the swirling primary

A. Guessab, Ecole Nationale Polytechnique d'Oran, Algérie, 31000. Départment de Genie Mécanique, Oran (e-mail: ahmed.guessab@enp-oran.dz).

A. Aris, Ecole Nationale Polytechnique d'Oran, Algérie, 31000. Départment de Génie Mécanique (e-mail: arisaek@yahoo.fr).

T. Benabdallah, Industrial Products Systems Innovation (PSIL), ENP, Oran, Algeria, 31000. Department of mechanical, (corresponding author: e-mail: tawfikbenabdallah@yahoo.fr, Phone: 00213555582016).

N. Chami, Industrial Products Systems Innovation (IPSIL.), ENPO, Algeria.

air flow. There are six small fuel inlets, each with a surface area of 0.14 cm^2. The secondary air is injected in the combustion chamber through six side air inlets each with an area of 2cm^2. The secondary air or dilution air is injected at 0.1 m from the fuel injector to control the flame temperature and NO_x emissions. The can-type combustor outlet has a rectangular shape with an area of 0.0150 m^2. In the present study, unstructured grid has been employed due to the complexity of geometry combustor. The 3-D modeling of the combustor has been done using the pre-processors Gambit (Figs. 3 and 4) [5]. A simplified 3-dimensional solid model has been built and used to generate the computational grid. The resulting solid model is shown in Figure 1. It was constructed from the dimensions shown in Figure 2. The mesh is generated by automatic method and its main quality specifications are summarized in Table I. The geometry is complex and consists of five separate solid bodies.

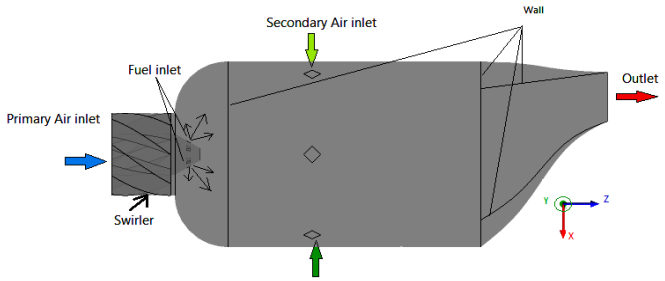


Fig. 1 Solid model of combustor flow domain.

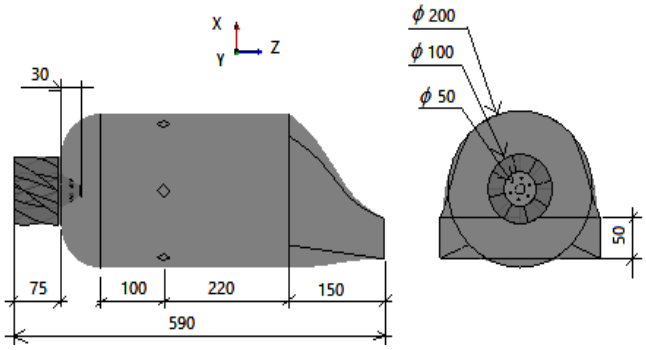


Fig. 2 Dimensions of combustor flow domain. All dimensions are in mm.

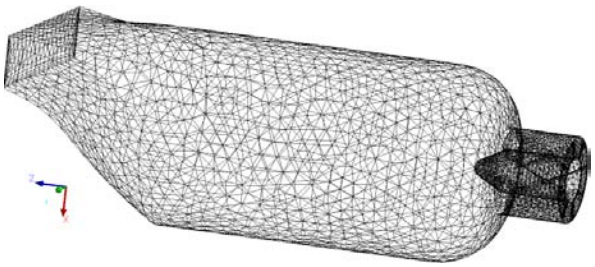


Fig. 3 Combustor computational mesh.

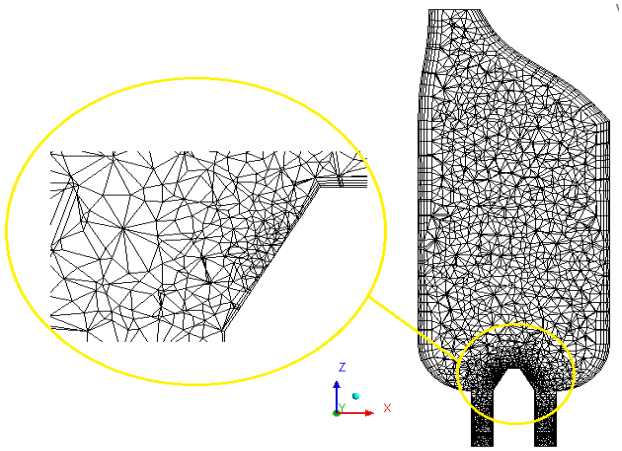


Fig. 4 Mech geometry of gas turbine can combustor (midaxial vertical plane).

Table I: Mesh statistics

Number of Nodes:	31433
Number of Elements:	106651
Tetrahedra:	74189
Pyramids:	1989
Wedges:	30473
Extents:	
min x, max x:	-0.11314 [m], 0.11314 [m]
min y, max y:	-0.125 [m], 0.125 [m]
min z, max z:	-0.138145 [m], 0.45 [m]
Max Edge Length Ratio:	21.1566
Volume:	0.0186368 [m^3]

The different boundary conditions applied for this numerical simulation of gas turbine can-type combustion chamber are: The velocity boundary condition is used to define the flow velocity along with all relevant scalar properties of the flow at the flow inlet. The velocity magnitude and direction, the velocity components or the velocity magnitude normal to the boundary specifies flow inlet velocity. These values were taken from rig tests simulating combustor conditions following a gas turbine engine. The outlet boundary condition is used to model the flow at exit where the details of the flow velocity and pressure are not known prior to solving the problem. As these variables were not known for the case under study, the outlet boundary condition was applied for the combustor exit. When this condition was specified, the code extrapolated the required information from the interior. In the present study, Outlets were set as pressure outlets with 0 Pa (gauge pressure) for combustor outlet, mass flows were calculated accordingly. In any flow, Reynolds number of the flow becomes very low and turbulent fluctuations are damped considerably, near the walls. The laminar viscosity starts to play a significant role. In the present case, walls were assumed to be adiabatic with no-slip condition. The standard wall functions were used to calculate the variables at the near wallcells and the corresponding quantities on the wall. The swirl components were directly introduced in the swirler air flow using ANSY Fluent 14 settings. The finite volume method and the second-order upwind method were used to solve the governing equations. The convergence criteria were set to 10^{-4} for the mass, momentum, turbulent kinetic energy and the dissipation rate of

ISBN: 978-1-61804-253-8

the turbulent kinetic energy and the chemical species conservation equations. For the energy and the pollution equations, the convergence criteria were set to 10^{-6}. Table II lists the numerical conditions.

III. Government Equations

In this paper we will present a simulation of a combustion process inside a gas turbine combustor, showing the NO and CO_2 formation zones. For this simulation we've used the ANSYS FLUENT software. The ANSYS FLUENT 14 solver [6] is used to solve the governing equations for the turbulent non-premixed combustion flame. The flow model considered in the present investigation is based on 3-D situation, steady state and turbulence closure model for RANS. However, the utilized model for governing turbulence-chemistry interaction is the eddy dissipation concept (EDC). ANSYS Fluent CFD modeling is based on solving transport equations for continuity, momentum, energy and species conservation with the use of Reynolds decomposition method.

Table II: Inlet boundary conditions.

Primary air	• Flow regime: Subsonic; • The injection velocity is 10 m/s; • The injector diameter is 85 mm. •Heat transfer: Static temperature: 300K • The turbulence intensity is 10%; • Mass fraction of oxygen :Y_{O2}= 0.232
Fuel	• Flow Direction: Normal to Boundary condition; • Flow regime: Subsonic; • The injection velocity is 40 m/s; • The injector diameter is 4.2 mm. • The temperature is 300K; • The turbulence intensity is 10%; • The species mass fraction is: (See Table 3) • Thermal radiation: Local temperature;
Secondary air (dilution)	• The injection velocity is 6 m/s; • The injector diameter is 16 mm. • The temperature is 300K; • The turbulence intensity is 10%; • Mass fraction of oxygen :Y_{O2}= 0.232
Outlet	• Flow regime: Subsonic; • The relative pressure is 0 Pa; • Mass fraction of oxygen :Y_{O2}=0.232; • Thermal radiation: local temperature;
Wall	• Wall boundary condition was no slip; • Wall roughness was smooth; • Wall heat transfer was adiabatic; • Wall emissivity was 0.95 • Thermal Radiation: Opaque. • Diffuse fraction: 1.

Closure for the Reynolds stress terms in the government equations were achieved using the RNG k-ε turbulence model. The RNG k-ε model was derived using a rigorous statistical technique called renormalization group theory (RNG) [7]. It is similar in form to the standard k-ε model, but includes the effect of swirl on turbulence for swirling flows.

The influence of turbulence on the reaction rate is taken into account by employing the Magnussen model [8]. In this model, the rate reaction $R_{i',k}$ is given by smaller (i.e., limiting value) of the two expressions below:

$$\hat{R}_{i',k} = \upsilon'_{i,k}\, M_i A.\rho \frac{\varepsilon}{k}\left(\frac{Y_R}{\upsilon'_{R,k}.M_R}\right) \tag{1}$$

$$\hat{R}_{i',k} = \upsilon'_{i,k}\, M_i AB.\rho \frac{\varepsilon}{k}\left(\frac{\sum_P Y_P}{\sum_{j'}^{N} \upsilon''_{j',k}.M_{j'}}\right) \tag{2}$$

Where
Y_P represents the mass fraction of any product species
Y_R represents the mass fraction of a particular reactant
M represents the molecular mass of species
A an empirical constant equal to 4.0
B an empirical constant equal to 0.5

The EDM relates the rate of reaction to the rate of dissipation of the reactant and product containing eddies. (k/ε) represents the time scale of the turbulent eddies following the Eddy dissipation model (EDM) of Patankar [9]. The model is useful for prediction of premixed and diffusion problems as well as for partially premixed reacting flows. The Air and fuel within the internal combustor flow field was modeled as a reacting, variable composition mixture. The two-step of Westbrook and Dryer [10] was used to compute the rate of fuel oxidation.

$$2CH_4 + 0.22N_2 + 3(O_2 + 3.76N_2) \rightarrow 2CO + 4H_2O + 11.5N_2 \tag{3}$$

$$2CO + (O_2 + 3.6N_2) \rightarrow 2CO_2 + 3.76N_2 \tag{4}$$

The constants of Arrhenius equations (3) and (4), reactions, the value of the pre-exponential factor, temperature exponent and activation energy for each reaction and the exponent rate for each species are reported in Table III.

Table III: The constants of the Arrhenius equations.

Eqs.	A_k [s^{-1}]	E_k	β_k	γ_{CH4}	γ_{O2}	γ_{CO}
28	1.50×10^{7}	30	0	-0.3	1.3	__
29	$1.0\times10^{-0.75}$	40	0	0	0.25	1

IV. Zeldovitch Reaction and NO_X Formation

There are two mechanisms that create NO_x in gas turbine combustor:

- Thermal NO_x, which is the oxidation of atmospheric bound nitrogen in the combustion air.
- The conversion of fuel bound nitrogen into NO_x.

Thermal NO_x is found by a series of chemical reactions that involve the dissociation and subsequent reaction of oxygen and nitrogen to form NO_x, following the well-known Zeldovich mechanism [11]. The Zeldovich reaction is represented by:

$$N_2 + O \Rightarrow NO + N \quad (5)$$

$$N + O_2 \Rightarrow NO + O \quad (6)$$

Net reaction: $N_2 + O_2 \Rightarrow 2NO$ (7)

- Fuel-bound NO_x is usually of less importance for normal fuels. Molecular nitrogen present in some kinds of natural gas does not contribute significatly to NO_x, but my be important with some low Btu fuels.

V. Results and Discussion

A. The effect of swirl angles

We begin the study of different flow configuration of various swirl angles (35°, 40° and 45°) are compared in the primary chamber between a swirling angular jet (Primary air) and the non-swirling jet (fuel). Swirling air flow is a key feature in many types of combustors.

Tangential flow component is generated in an aerodynamics element called swirler (swirl generator). Such design in low-NO_x diffusion burners with staged gas. The swirler is a key burner design component that significantly influences the flow pattern in combustion chambers. This different configuration are altered to examine about the emission and to obtain effective combustor with less NO_x and CO_2 emission.

Figure 5 demonstrates the temperature for the can-type combustor with different swirler angles at axial distance. The temperature increases gradually due to the chemical reaction inside the main combustor. It is clear that after the location of primary chamber at z = 0.25m, the temperature diminishes due to the cooling effect of air from the secondary holes. Thus, it is distinct to show that the 45 degree swirler angle achieves better temperature in comparison with other swirler angles.

Depending upon the temperature the NO emission is determined by the Zeldovich mechanism. The main intention for introducing more air in the secondary chamber is to reduce the NO emission. Figure 6 demonstrates the NO mass fraction for the can-combustor with different swirler angles and axial distance. It is evident from the figure that after the location of primary chamber at Z=0.25m, the emission of NO diminishes because of the cooling effect of air. Hence, it can be stated that the 45 degree swirler angle attempts less NO emission due to low exit temperature.

Figure 7 demonstrates the CO_2 mass fraction for the can-combustor with different swirler angles and axial distance. The air at the secondary inlet is introduced to reduce the NO emission and also to mitigate CO_2 emission from natural gas. In Fig. 7, it is clearly predicted that CO_2 mitigated to the value of 10% due to the cooling effect of air from the secondary inlet by swirling velocity.

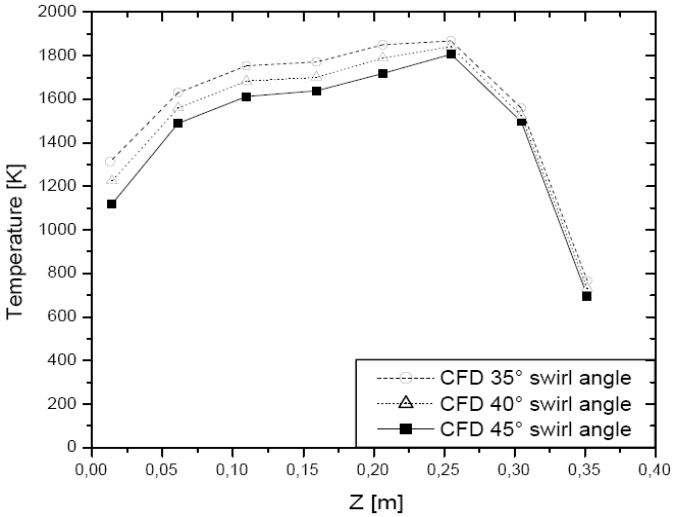


Fig. 5 Axial temperature for the can-combustor with different swirler angles

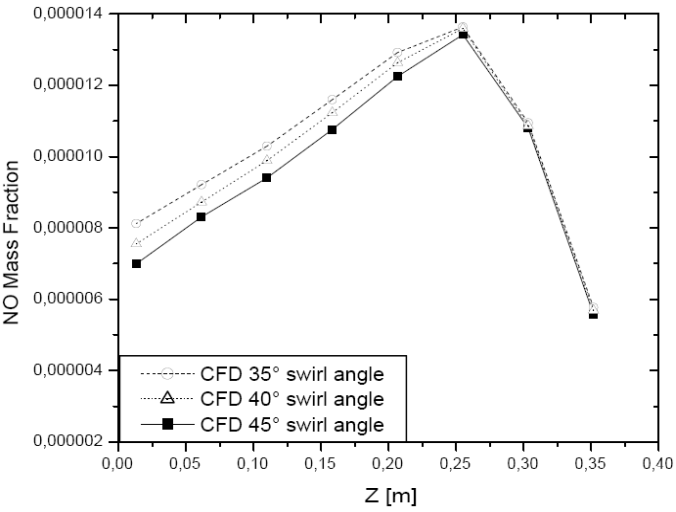


Fig. 6 Axial NO mass fraction for the can-combustor with different swirler angles

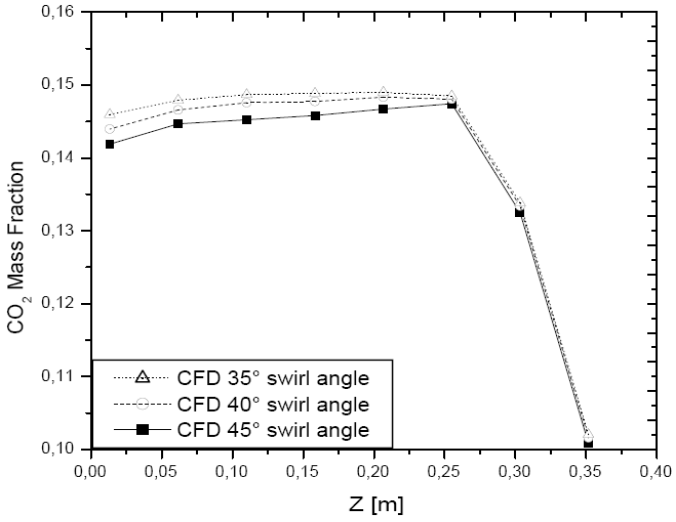


Fig.7 Axial CO_2 mass fraction for the can-combustor with different swirler angles

B. Effect of Fuels on species emission and temperature

In this study, methane is used as the reference fuel. Waste biogas is composed of Biogas 1 (%CH_4=90, %CO_2=10), Biogas 2 (%CH_4=75, %CO_2=25) and Biogas 3 (%CH_4=70, %CO_2=30), with a constant velocity inlet of fuel and air. In the current studies, the vane angle was assumed to be about 45° and was kept constant for all the simulation. The simulations were validated in terms of temperature and emission of pollutants NO and CO_2. The temperature distribution along the central axis of combustor is shown in Fig. 8. As seen in the figure, due to more fuel burned under the increase percentage of CO_2, the outlet temperature decrease as the CO_2 increases. In the region from the fuel injector to z = 0.1m, the temperature at the central axis of combustor has little change for all fuels. However, in the region from 0.1 m to combustor outlet, the temperature decrease as the percentage of CO_2 increases. It can also be seen from Fig. 8 due to the air from the secondary holes takes more heat to reach the equilibrium temperature and the combustion process the mixing of dilution air and fuel gas is useful to improve the uniformity of combustor outlet temperature distribution. Figure 8 demonstrates the temperature for the can-type combustor with different fuel properties and axial distance. The temperature increases gradually due to the chemical reaction inside the main combustor. It is clear that after the location of primary chamber at z = 0.25m, the temperature diminishes due to the cooling effect of air from the secondary holes. Thus, it is distinct to show that for the biogas 3, achieves better temperature in comparison with other fuel. The peak gas temperature is located in the secondary zone (z = 0.25m) for methane and biogas 1. Then the peak gas temperature is located in the primary zone for Biogas 2 and biogas 3 (z = 0.1m). However, due to the dilution of burned mixture gas with the air, the gas mixture temperature is lower in primary zone. Adding about 10% of CO_2 to methane lowers the adiabatic temperature of combustion by about 10K, and increasing the ratio of CO_2 to 30% by about 45K. The highest temperature in the combustion chamber was recorded for the burning of a mixture of methane. Increasing the amount of carbon dioxide in the fuel leads to lowering of the maximum combustion temperature. Figure 9 demonstrates the NO mass fraction for the can-type combustor with different fuels at axial distance. The highest values of distribution of the mass fraction of NO in the combustor chamber were measured for fuels with the highest combustion temperature. NO concentrations, however, are highest in the slow-moving, reverse-flowing fluid near the combustor centerline. The fluid here has had sufficient time to allow N_2 to react with O_2 and form NO. An analysis of CO_2 distribution in the combustion chamber in the axis the combustor was carried for methane pure and mixture of methane and CO_2 (Fig. 10). The proper selection of a swirler is required to reduce the emission which can be concluded from the emission of NO and CO_2. The air at the secondary inlet is introduced to reduce the NO emission and also to mitigate CO_2 emission from natural gas.

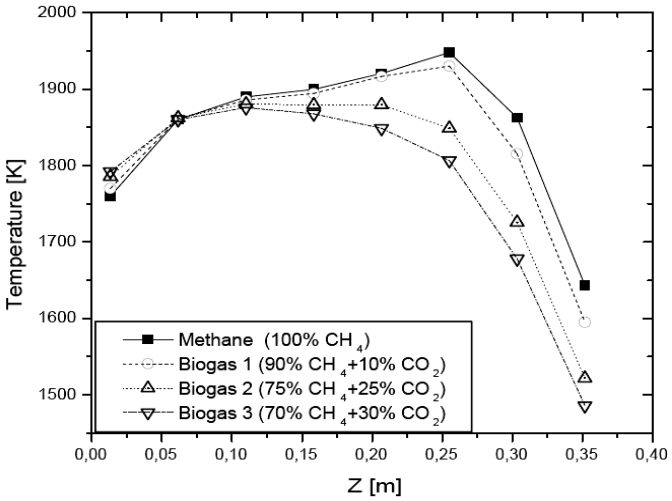


Fig. 8 Distribution of temperature along the center line of combustor.

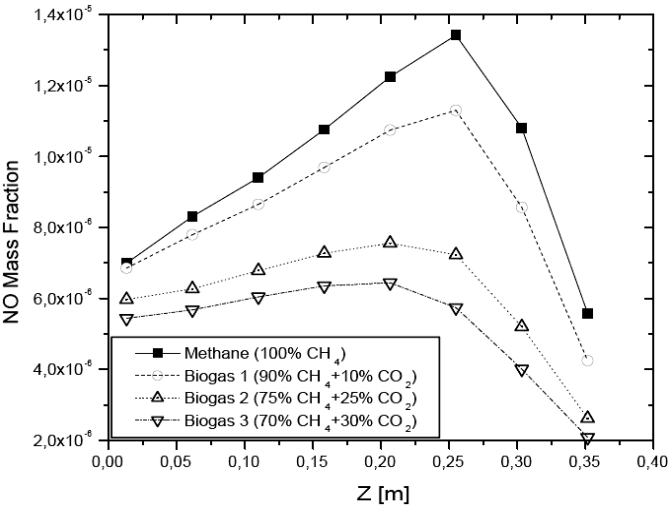


Fig. 9 Distribution of NO mass fraction along the center line of combustor.

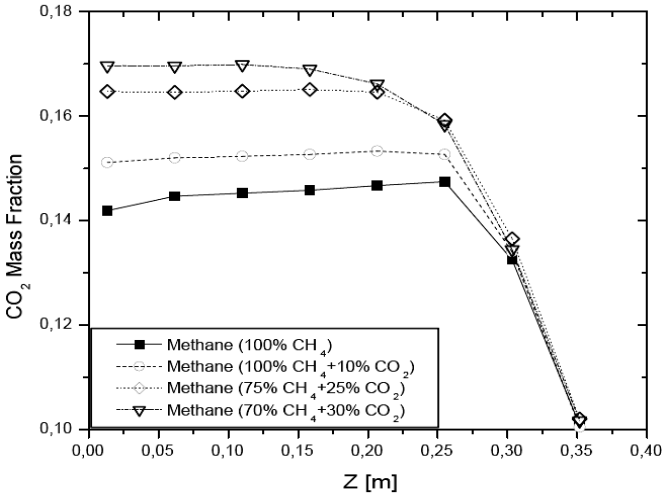


Fig. 10 Distribution of CO_2 mass fraction along the center line of combustor.

Figure 11 presents a set of measurements of toxic compounds or mixtures of natural gas (100% CH4) and biogas 1, 2 and 3 burnt in strong swirl flows. The values measured for NO correspond to those taken 30mm from the rectangular shape outlet of the combustor which shows that the whole of NO emissions is produced no higher than the flame height indicated and that in the flame no reburning of nitric oxides takes place.

The emission of CO measured is much smaller, which points to the process of their afterburning in hot exhaust gases in the upper parts of the flame.

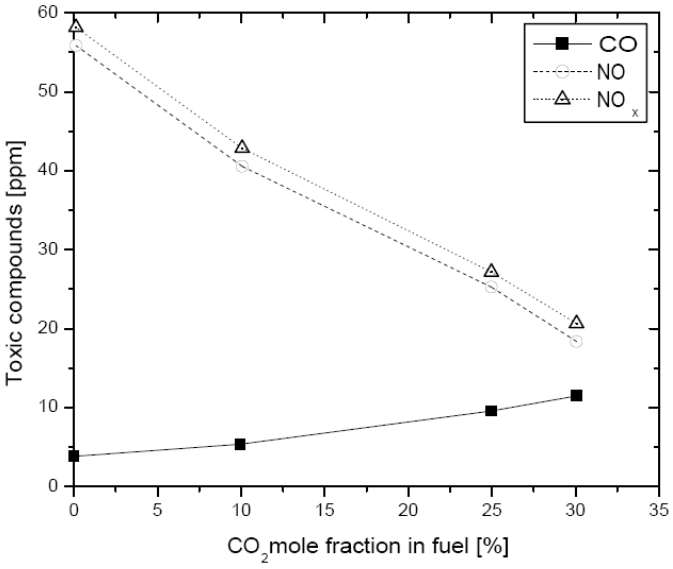


Fig. 11 Emission of NO, NO_x and CO vs. CO2 content in Combustion mixture

V. Conclusion

Combustion simulations have been performed on gas turbine can-type combustor with the use of the commercial CFD software ANSYS-FLUENT. The effect of premixing CO_2 and methane on NO emissions was investigated. Due to previous experimental problems with stability in the actual burner when burning methane, biogas was chosen as fuel.

For modeling of the gas turbine can-type combustor, a 3D un-structured grid consisting of about 106651cells were created in GAMBIT. The grid was imported into ANSYS-FLUENT where simulations of four different non-premixed air-fuel ratios. For the modeling of turbulence the RNG k-ε Model was employed, while the Eddy Dissipation Model was used for modeling combustion.

The results from the simulations proved satisfactory when comparing with previous work (Fuel= 100% Methane) done on the same combustor. The conclusion regarding the analysis of the process of combustion of mixtures of natural gas with CO_2.

The results of this investigation have also clearly demonstrated that it is possible to use such fuels in combustion systems with swirl burners. Carbon dioxide contain in the fuel leads to the lowering of the temperature of the flame, the effect of which is a reduction in emissions of NO and CO. Numerical investigation on Can-type combustion chamber shows that biogas 3 (70% CH_4 + 30 % CO_2) is giving less NO emission as the temperature at the exit of combustion chamber is less as compared to gas natural, biogas 1 and biogas 2. Temperature profiles shows increment at reaction zone due to burning of air-ful mixture and decrement in temperature at the downstream of secondary inlet holes due to supply of more to dilute the combustion mixture.

The results from the parametric studies indicate that the calculation of NO emission serves to develop low emission combustor.

Acknowledgment

The authors would like to thank University of ENPO and the "SONATRACH" company (LNG group-Algeria) for the support regarding this research Project.

References

[1] H. Pathan, K. Partel, and V. Tadvi, "Numerical investigation of the combustion of methane air mixture in gas turbine can-type combustion chamber," *International Journal of Scientific & Engineering Research*, Vol. 3, No. 10, 2012, pp. 1-7.

[2] P. Koutmos and J. J. McGuirk, "Isothermal flow in a gas turbine combustor–a benchmark experimental study," *Experiments in Fluids*, Vol. 7, 1989, pp. 344-354.

[3] C. Ghenai, "Combustion of syngas fuel in gas turbine can combustor," *Advances in Mechanical Engineering*, Vol. 1, 2010, pp. 1-13.

[4] Y. A. Eldrainy, J. Jeffrie, and M. Jaafar, "Prediction of the flow inside a Micro Gas Turbine Combustor," *Journal of Mechanical*, vol. 25, 2008, pp. 50-63.

[5] ANSYS Gambit 2.4, http://www.ansys.com.

[6] ANSYS FLUENT v. 12. , Theory Guide: Release 12.0, Last modified , April (2009). http://www.ansys.com.

[7] Franz D. (2008), An Introduction to the Theory of Fluid Flows, Fluid Mechanics, Springer.

[8] B.F. Magnussen, On the Structure of Turbulence and a Generalized Eddy Dissipation Concept for Chemical Reaction in Turbulent Flow. Nineteenth AIAA Meeting, St. Louis, (1981).

[9] Patankar S. V., "Numerical heat transfer and fluid flow", McGraw-Hill, New York, USA, 1983.

[10] Westbrook, C. K., Dryer, F. L., Simplified Reaction Mechanisms for the Oxidation of Hydrocarbon Fuels in Flames, Combustion Sciences and Technologies, 27(1981), 1-2, pp. 31-43.

[11] Y. B. Zeldovich, "The oxidation of nitrogen in combustion explosions," Acta Physicochim. URSS **21**, 577-1946.

Dr. Guessab Ahmed was born in Algeria, Planteurs City, Oran, in 1974. He received his Bachelor degree in mathematics from Ibn Badiss School, Oran. Engineer degree in Mechanical Engineering from Oran University Algeria (USTOMB) in 2002, and he has Magister Degrees, in energetic from mechanical institute Oran University (USTOMB) in 2004. He was a teacher of some disciplines in National Polytechnic School from 2005 to present day. He focuses his research interests on the turbulent flow, models of turbulent combustion and combustion CFD simulation.

Dr. Aris Abdelkader is an Assistant Professor of Mechanical Engineering at the National Polytechnic School of Oran, Algeria. He received his Ph.D in Mechanical Engineering from Oran University, USTO.MB (Algeria) in 2006. His researches focus on numerical and experimental combustion phenomenon (internal combustion engine, burners and furnaces), heat transfer and renewable energy. He is responsible for the team's energy specialty in engineering school (ENP.Oran)

Dr. Tawfik Benabdallah is a lecturer and head of Industrial Products & Systems Innovation Laboratory (IPSIL) and Involved in Trans Euro Mediterranean Projects and in many international Review committees.

Chami Nadir Post Graduate PHD Student, Electronic Engineering Specialist involved in IPSIL activities.

Effect of processing conditions on the mechanical properties of Polylactic acid/clay composites

Fares D. Alsewailem, Sushant Agarwal, Man Chio Tang, and Rakesh K. Gupta

***Abstract*—**In this research, composites of polylactic acid (PLA) filled with nano-clay were prepared, and their mechanical properties were measured. The aim was to investigate effect of processing conditions on tensile and impact strengths of the prepared composites. Two methods were used to study such effect. first, samples of PLA/clay were prepared by extrusion followed by injection molding, and in the other method samples of PLA/clay were prepared only by injection molding. It was found that mechanical properties slightly influenced by method of processing, i.e. one step processing by injection molding is preferred in order to have less degraded PLA/clay composites.

***Keywords*—** Clay, mechanical properties, Polylactic acid.

I. Introduction

Polylactic acid (PLA) is a biodegradable polymer which is generally obtained from renewable agricultural products. Although this bio-plastic was originally used for medical applications, such as absorbable sutures, current-day applications are in food packaging where optical clarity, flexibility and gas barrier properties are important. Key advantages of this polymer include biodegradability and a low melt-processing temperature. However, the plastic is inherently brittle, with a low elongation-to-break, and it has poor heat stability, which ultimately hinder its potential to be used in advanced applications such as in automotive. Therefore, improving mechanical properties of PLA is a noble objective in order to allow for PLA use in structural and other advanced applications.

As with all polymers, PLA properties depend on molecular weight, chain flexibility, presence of polar groups, amount, shape and size of crystals and chain orientation in the amorphous and crystalline regions. Most of the reported studies on PLA modified with inorganic fillers such as clay particles had been focusing on thermal stability and degradation whereas fewer studies had reported mechanical properties of such composites [1-4].

In this research we are investigating effect of processing conditions such as equipment used to prepare PLA/clay samples on the tensile and impact strengths of the composites.

This work was supported by King Abdulaziz city for science and technology (KACST).

F. D. Alsewailem is with KACST, Riyadh 11442, Saudi Arabia (corresponding author, phone: +96611-481-3522; fax: +96611-481-3670; e-mail: fsewailm@kacst.edu.sa).

S. Agarwal is with department of chemical engineering, West Virginia University, Morgantown, WV 26506 USA (e-mail: author@lamar. colostate.edu).

M. C. Tang is with department of chemical engineering, West Virginia University, Morgantown, WV 26506 USA (e-mail: author@lamar. colostate.edu).

R. K. Gupta is with department of chemical engineering, West Virginia University, Morgantown, WV 26506 USA (e-mail: author@lamar. colostate.edu).

II. Procedure

A. Materials used

PLA used in this research was Ingeo 2003D (high MW) and Ingeo 6751D (low MW) from Natureworks. These resins were used as the matrix material, while Cloisite 30B nano-clay from Southern Clay Products was used as the nano-fillers.

B. Sample preparation

Composites were made by melting the polymer and mixing it with nanoclay in a Dynisco mini extruder (LME) at 200^{o}C and 20 rpm for 3-10 min. A Dynisco mini molder (LMM) was then used to mold the compounded resin into tensile and impact samples according to the dimensions of ASTM D 1708 and ASTM D 256 standards for tensile and impact strength testing respectively.

C. Mechanical tastings

A Tinius Olsen pendulum impact tester (IT 504), with a total capacity of 7 J of energy, was used to measure notched Izod impact strength of the nanocomposites. Tensile strength measurements were carried out using a Hounsfield universal testing machine, model H5KS, using a 5 KN load cell.

D. Morphology of the composites

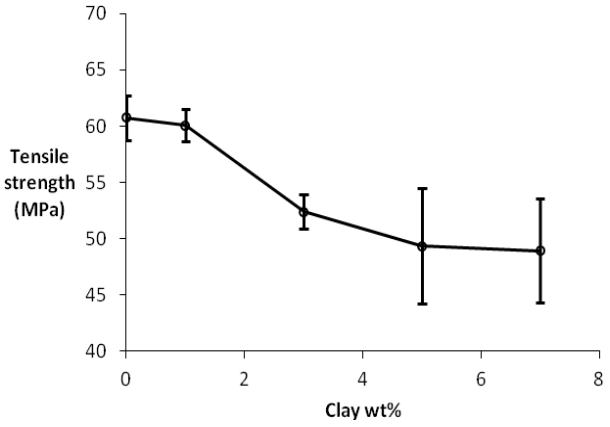
The fracture surfaces of the Izod samples were examined by the scanning electron microscope, SEM (FEI, model NNL 200).

III. Results

Figures 1-4 show the mechanical behavior of the prepared samples of PLA2003D and PLA6751D compounded with 1-7

ISBN: 978-1-61804-253-8

wt% of Cloisite 30B**.** Note that PLA2003D has a higher molecular weight as compared to PLA 6751D, and this accounts for the difference in the tensile and impact properties of the pure resins. It can be seen from the figures that generally impact strength of the composites increases upon increasing clay loading especially at higher loading of 7 wt%. On the contrary, tensile strength of the composites decreases upon increasing clay loading. Some of the samples as shown in Figures 3 and 4 were prepared directly by the mini molder (LMM) without mixing in the mini extruder (LME). This was done to investigate the effect of shear and heat degradation of PLA on the mechanical properties. Figures 3 and 4 show that samples prepared only by LMM have slightly higher values of Izod impact strength in comparison with those prepared by LME followed by LMM as shown in figures 1 and 2. Figures 5 and 6 show how the Izod specimens appear depending on the preparation method. The bars prepared by LME followed by LMM appeared somewhat dark in comparison with those prepared only by LMM. This may be an indication of degradation of the materials because prolong time of exposing to shearing and heat can cause degradation of polymeric material.



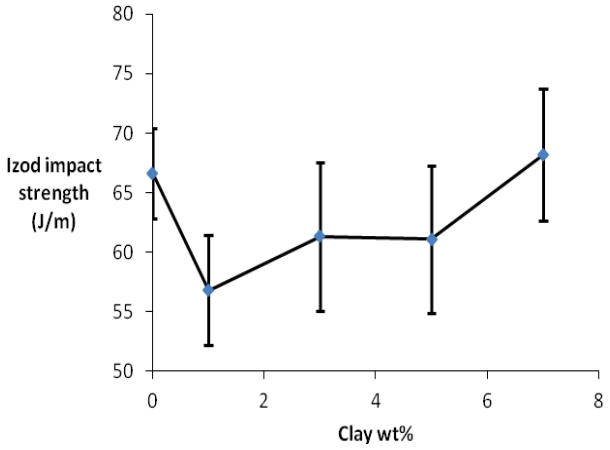
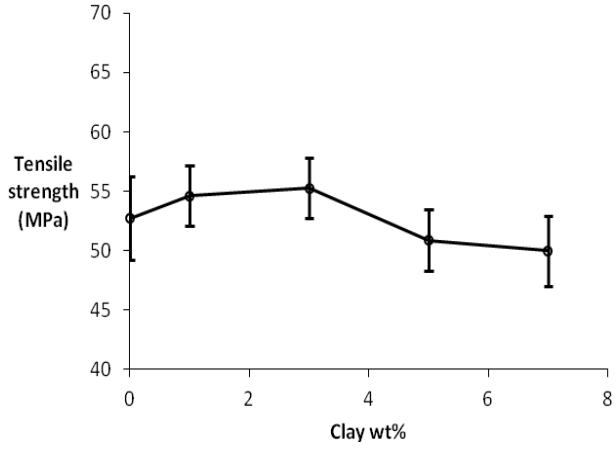


Figure 1 Tensile and impact strength of PLA2003D/ clay composites.



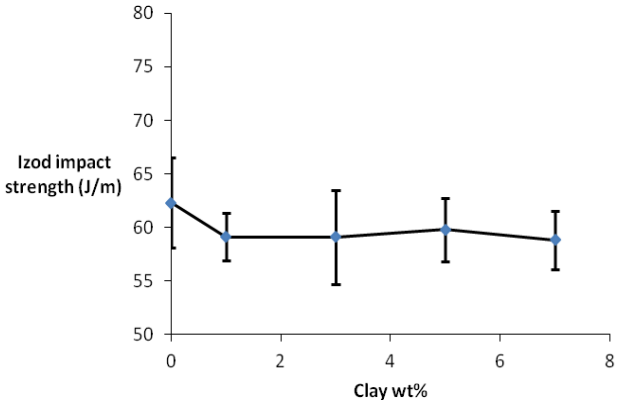


Figure 2 Tensile and impact strength of PLA6751D/clay composites.

Figures 7 and 8 show SEM of fractured surface of PLA/clay composites. At relatively lower loading of clay, i.e. less than 5 wt%, no dispersed particles were observed at a magnification of 15000x which may indicate that clay particles may be observed at much higher magnification where the scale of the graphs can be in submicron or nano size. At higher loading of clay, i.e. up to 5 wt%, one can see clay particles clustering around. A noteworthy case is the smooth morphology at relatively higher clay content, i.e. 5wt%, when the mixing in LME was skipped.

ISBN: 978-1-61804-253-8

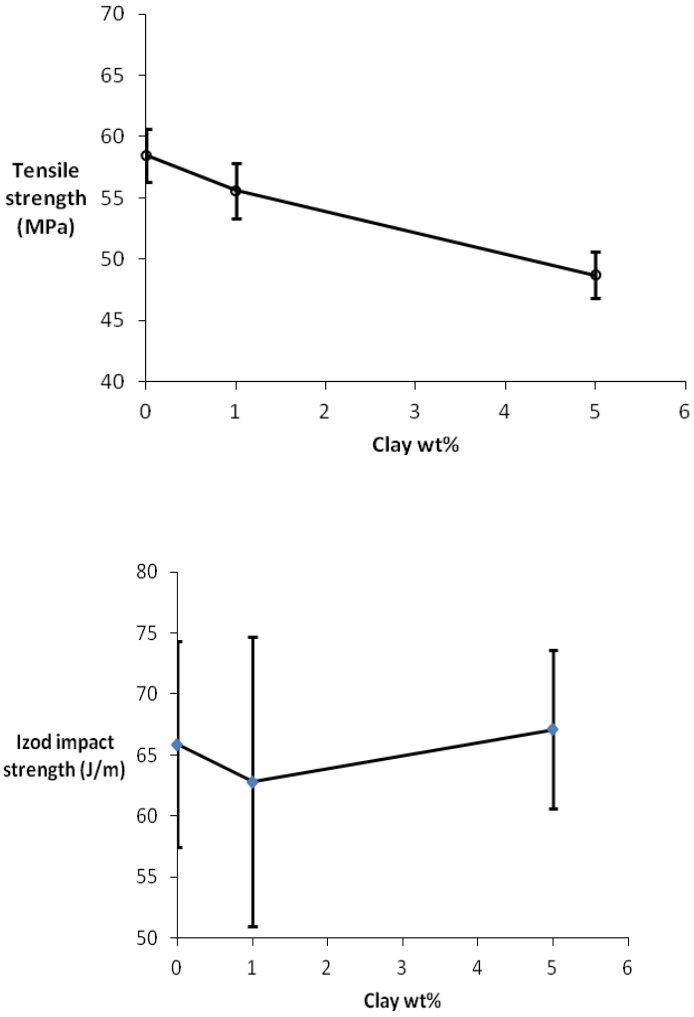


Figure 3 Tensile and impact strength of PLA (2003D)/ clay composites prepared by LMM only.

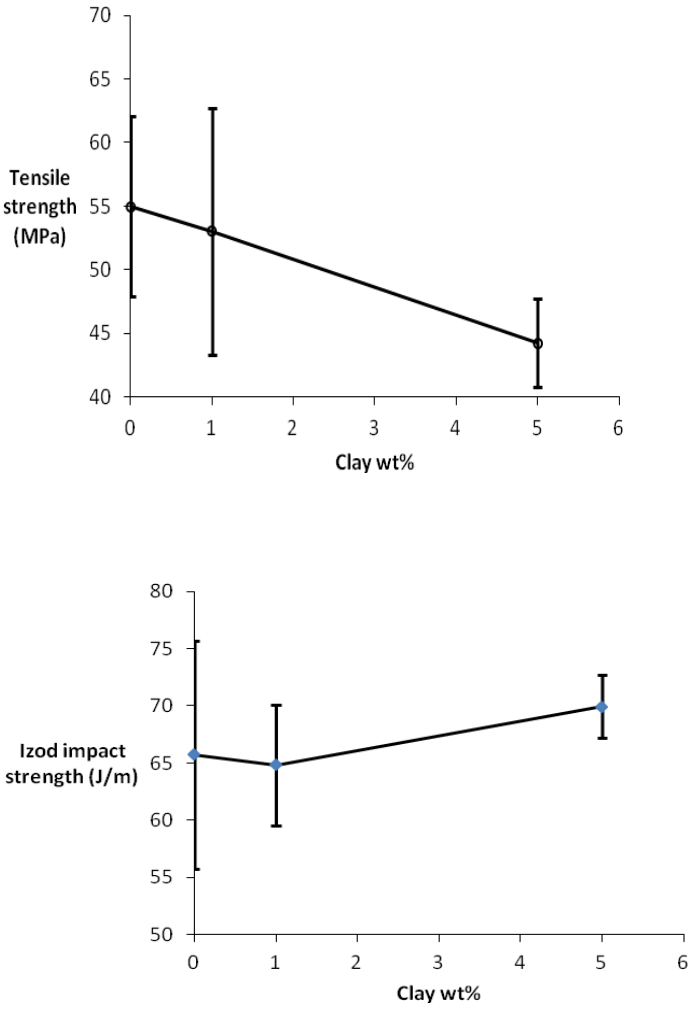


Figure 4 Tensile and impact strength of PLA (6751D)/ clay (Cloisite 30B) prepared by LMM only.

ISBN: 978-1-61804-253-8

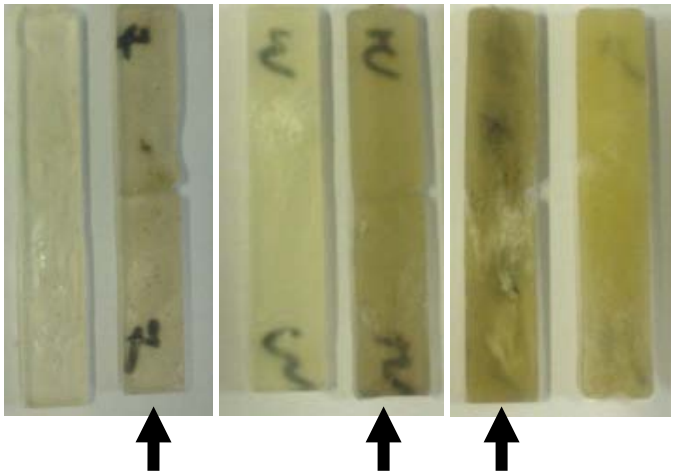


Figure 5 Appearance of Izod bars of PLA 6752D/clay composites. Arrows indicate samples prepared by LME plus LMM.

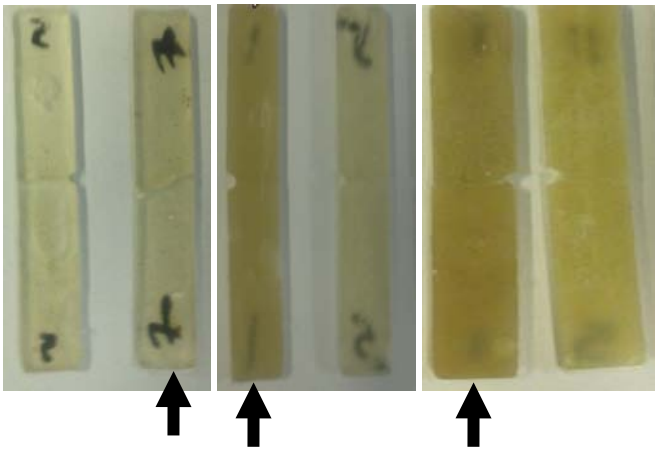


Figure 6 Appearance of Izod bars of PLA 2003D/clay composites. Arrows indicate samples prepared by LME plus LMM.

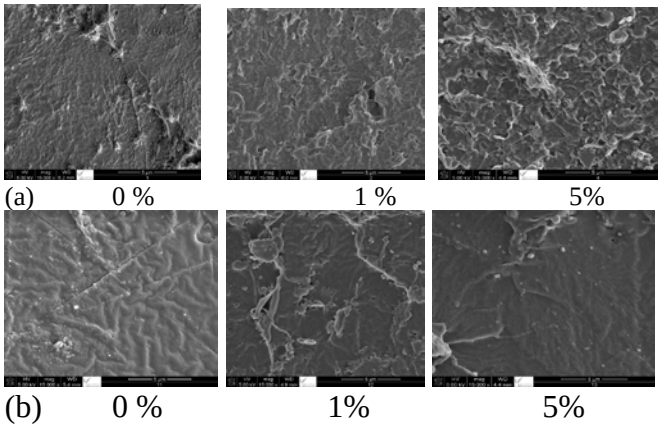


Figure 7 SEM micrographs of Izod samples of PLA 2003D with various weight percents of clay. (a) processed by LME followed by LMM. (b) processed by LMM only.

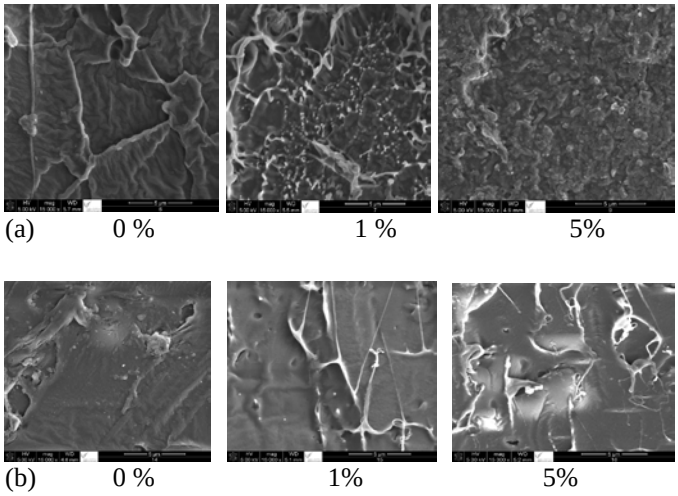


Figure 8 SEM micrographs of Izod samples of PLA 6752D with various weight percents of clay. (a) processed by LME followed by LMM. (b) processed by LMM only.

IV. Conclusion

PLA/clay composites prepared by only one step, i.e. injection molding were more likely to be less degraded with better mechanical properties compared to those prepared by extrusion followed by injection molding.

Acknowledgment

The authors thank KACST for financial support of this research.

References

[1] H. Ebadi-Dehaghani, H. A. Khonakdar, M. Barikani, S. H. Jafari, "Experimental and theoretical analyses of mechanical properties of PP/PLA/clay nanocomposites," *Composites: Part B*, vol. 69, pp. 133–144, 2015.

[2] K. Fukushima, D. Tabuani, M. Arena, M. Gennari, G. Camino," Effect of clay type and loading on thermal, mechanical properties andbiodegradation of poly(lactic acid) nanocomposites," Reactive & Functional Polymers, vol. 73, pp.540–549, 2013.

[3] S. M. Lai, S. H. Wub, G. G. Lin, T. M. Don," Unusual mechanical properties of melt-blended poly(lactic acid) (PLA)/clay nanocomposites," European Polymer Journal, vol. 52, pp.193–206, 2014.

[4] M. Yourdkhani, T. Mousavand, N. Chapleau, P. Hubert," Thermal, oxygen barrier and mechanical properties of polylactide–organoclay nanocomposites," Composites Science and Technology, vol. 82, pp.47–53, 2013.

Estimation of Heat Loss from a Cylindrical Cavity Receiver Based on Simultaneous Energy and Exergy Analyses

Vahid Madadi [1], Touraj Tavakoli [2,*], Amir Rahimi [3]

***Abstract*—** This study undertakes the experimental and theoretical investigation of heat losses from a cylindrical cavity receiver employed in a solar parabolic dish collector. Simultaneous energy and exergy equations is used to thermal performance analysis of system. The effects of wind speed and its direction on convection loss has also been investigated. The effects of operational parameter such as Heat Transfer Fluid (HTF) mass flow rate, wind speed and structural parameters such as receiver geometry and inclination are investigated. The portion of radiative heat loss is less than 10%. An empirical and simplified correlation for estimating of dimensionless convective heat transfer coefficient in terms of Re number and average receiver wall temperature is proposed. This correlation is applicable for wind speed range of 0.1 to 10 m/s. Also the proposed correlation for Nu number is validated by using of experimental data obtained through the experiments carried out with a conical receiver with two aperture diameters. The R2 and RMSE parameters were calculated and results show that there is a good agreement between predicted results and experimental data. The R2 is greater than 0.95 and the RMSE parameters is less than 4 in this analysis.

***Keywords*—** Energy analysis; Exergy analysis; Receiver; Radiative heat loss; Convective heat loss; Nusselt number correlation

I. Introduction

In many situations it is desirable to provide energy at temperatures higher than those possible with flat-plate collectors. For this purpose, direct solar radiation can be concentrated by a solar collector system. The solar energy concentrating technologies may be classified as: linear Fresnel reflectors (LFRs), parabolic trough collectors (PTCs) and parabolic dish collectors (PDCs). Among solar thermal technologies, PDC systems have demonstrated the highest efficiency, producing a concentration ratio of more than 3000 and operating at temperatures of 750 ºC at annual efficiencies of 23% [1, 2]. PDC systems, in general, comprise a parabolic dish concentrator and a solar receiver located at the focus of the dish. As the solar receiver plays a role of transferring the solar radiation to the Heat Transfer Fluid (HTF), and heat loss of the solar receiver can significantly reduce the efficiency and consequently the cost effectiveness of the system, it is important to assess and subsequently improve its thermal performance and must be well designed to achieve high temperatures with minimal heat losses.

The thermal losses of a solar cavity receiver include the conduction heat loss through the cavity insulation and convection and radiation heat losses from the cavity to the ambient air. The estimation of conduction and radiation heat losses is more simplify than convection heat loss. The conduction heat loss is dependent on the receiver temperature and the thermal properties of insulation material, the radiation heat loss is dependent on the receiver wall temperature, the shape factors and emissivity of the receiver walls, while, the determination of convection heat loss due to the complexity of the temperature and velocity fields in and around the solar cavities, is rather difficult. There are too many factors that influence the convection heat loss of cavity receivers; the inclination and the geometry of cavity, the external wind conditions around the cavity and the air temperature within the cavity. The conduction and radiation heat losses from the receiver to the ambient air can be determined analytically [3, 4]. So far, the literature survey shows that most of studies on thermal performance of the solar cavity receiver focus on the heat loss and a few of them have engaged in exploring the mechanisms of convective heat loss for cavity receivers [5-8].

Many experimental and analytical studies have been done on convection heat loss from cylindrical receivers. A model was proposed by Koenig and Marvin which was appropriate for receiver temperature between 550 up to 900 ºC [9]. Stine and McDonald proposed an explicit model accounts for the combined effects of operating temperature, tilt angles and aperture size [2]. Leibfried and Ortjohann modified the Stine and McDonald model. In their study, the effect of wind on receiver losses for an upward-facing cavity is investigated. Results showed that the heat loss from the receiver is a function of the wind direction and ventilation decreases the

This work was supported in part by the U.S. Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials.

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

loss by more than 11%. For small wind speeds, the dependence on wind for sideward and downward-facing cavities is larger than for upward-facing cavities. For typical windy conditions, both types of cavity receivers are expected to have the same magnitude of convection heat loss [7]. Lovegrove et al. studied the heat losses from cavity receivers which are employed in solar parabolic dishes. Researchers developed a correlation that can reliably predict natural convection and proposed a model to account for the combined effect of the cavity geometrical parameters and the inclination by using the L_e as the characteristic length [10]. A numerical investigation of natural and combined convection heat loss from cavity receivers was done by Paitoonsurikarn and Lovegrove. They proposed a new correlation for prediction of heat transfer coefficients. In their study, the ensemble cavity length, L_s, was modified to include the aperture geometry [11]. Paitoonsurikarn et al. accomplished a numerical investigation of natural and combined convection heat loss from cylindrical cavity receivers employed in solar parabolic dishes. In this study, a parametric study of several relevant parameters was carried out and the previously proposed correlation model in Paitoonsurikarn and Lovegrove [11] has been modified. In modified model, the variation of parameters was implemented and moreover, a correlation based on the modified Stine and McDonald model was developed [12]. In another research, the interaction between the wind and the dish structure was studied by Paitoonsurikarn and Lovegrove.

The results of numerical simulations showed that, the magnitude and the direction of the wind can greatly affect the amount of convection heat loss. In addition, an improved version of correlation was presented based on the results of previous works, [11, 12] and numerical simulation results [13, 14]. An experimental and numerical study of the steady state heat losses occurring from a downward-facing cylindrical cavity receiver was carried out by Prakash et al. . In this study the effects of external wind at two different velocities in two directions was investigated. The results showed that the convection heat loss from head-on wind is higher than the side-on wind, which have conflict to previous reported results by [14, 15].

In order to analyze the convection heat loss from the solar receivers, two approaches are employed: isothermal receiver wall conditions and iso-flux receiver wall condition. In this study, first approach is employed in thermal analysis of system and assume that the receiver wall is isothermal. In the experiments related to the first approach finding the average receiver wall temperature is complicated and in some cases is unpractical due to receiver geometry and HTF circulating path through the receiver.

In our previous work, an energy and exergy analysis was carried out for the system under study [16] and in this study, a new technique is applied to find the average receiver wall temperature. Here, the average receiver wall temperature is found by simultaneous applying of first and second laws of thermodynamic (energy and exergy analysis). The heat loss from a cylindrical receiver in presence of wind, when the wind direction is parallel to aperture plane, is investigated. The aim of this study is to provide a simplified and applicable correlation for the Nusselt number to estimate the heat loss from solar receivers with various geometries in parabolic dish systems. A correlation for Nusselt number as a function of Reynolds number and average receiver wall temperature is proposed. For validating the correlation, a conical receiver is studied. Results show that, there is a good agreement between predicted outcomes and experimental data.

II. Experiments and Methods

The system under study consists of a parabolic dish concentrator with a cylindrical receiver which is employed in the center of dish. Two supporting adjustable metal arms are installed on the dish frame to adjust the receiver on the focal center of the dish. A photo of dish with cylindrical receiver is shown in Fig. 1. The dish aperture diameter and focal length are 2.88 and 1.5 m respectively. The sunlight tracking is manual. Three receiver aperture diameters, 0.115, 0.14 and 0.2 m were applied in experiments. The receiver height is 0.4 m and HTF moves in a spiral path with 0.03 m gap space. The mirrors are stuck on the metal surface and the entire system is installed on a concrete foundation. Through experiments, two parameters of HTF is measured, HTF temperature and mass flow rate. The HTF inlet and outlet temperatures are measured by two PT100 RTD thermocouples, and the temperatures are shown in a digital monitor which is installed in the back of the dish. The HTF inlet temperature range was 285 to 325 K. Also, a flow meter is used to measure the HTF mass flow rate and a range between 0.007 up to 0.5 kg/s were tested. The ambient air temperature and velocity are measured by Lutron ANEMOMETER/HUMIDITY METER Model AM-4205A device. The amount of solar radiation is measured by two devices for more accuracy, TES-1333/TES-1333R Solar Power Meter and TES-132 Solar Power Meter.

III. 3. Energy and Exergy Analysis

The direction of wind is an essential and important factor and plays a large role in determining the heat loss from solar receivers. In this analysis, some of the experiments were so arranged that in which the wind direction was normal to dish focal line. Wind direction is schematically shown in Fig. 1. For steady state conditions, the energy balance for the cylindrical receiver (control volume shown in Fig. 1) can be written as following equation:

$$\eta_o I_b A_c + \dot{m} c_P \left(T_{in} - T_{out}\right) - \dot{Q}_{l,conv.} - \dot{Q}_{l,rad.} = 0 \qquad (1)$$

where, η_o is the optical efficiency which is defined as the amount of reflected solar radiation from the concentrator to the receiver and in this study is considered as 0.75. In Eq. (1), I_b is global solar radiation, A_c is concentrator aperture area, $\dot{Q}_{l,conv.}$ and $\dot{Q}_{l,rad.}$ are the rates of heat which are lost by convection and radiation mechanisms respectively.

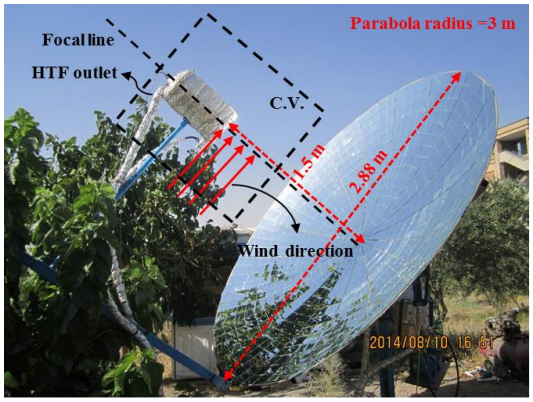


Figure 1: The photo of dish with cylindrical receiver

For isothermal wall condition, the receiver emissivity is constant and the entire inner surface of receiver can be considered as single surface and the surrounding as second surface. The rate of radiation heat loss can be estimated from the following equation [17]:

$$\dot{Q}_{l,rad.} = \varepsilon_{eff}\sigma A_{r,\mathrm{ap}}\left(T_w^4 - T_a^4\right) \tag{2}$$

where, σ is Stefan–Boltzmann constant, $A_{r,ap}$ is receiver aperture area, T_w and T_a are receiver wall and ambient temperatures respectively. The $\varepsilon_{eff.}$ is effective emissivity and is based on the receiver total surface area and is given by:

$$\varepsilon_{eff} = \frac{1}{1+\left(\frac{1-\varepsilon}{\varepsilon}\right)\frac{A_{r,ap}}{A_{r,w}}} \tag{3}$$

By combining Eqs. (1) and (2), the energy balance for receiver can be rewritten as below:

$$\eta_o I_b A_c + \dot{m}c_P\left(T_{in} - T_{out}\right) - U_l A_{r,ap}\left(T_w - T_a\right) - \varepsilon_{eff}\sigma A_{r,\mathrm{ap}}\left(T_w^4 - T_a^4\right) = 0 \tag{4}$$

Here, it is assumed that, the receiver wall is isothermal and all elements of receiver inner surface have the same temperature. But in experiments, temperature measurement of all parts of receiver surface is approximately impossible and impractical. In Eq. (4), two unknown parameters are exist: receiver wall temperature, T_w and convection overall heat transfer coefficient, U_l. So, to define the total rate of heat loss from receiver, we need another equation. By applying the second thermodynamic law at steady state condition for selected C.V. shown in Fig. 1, the rate of total convection heat loss can be defined. The steady state exergy balance for the C.V. is given by:

$$\sum\dot{E}_{in} - \sum\dot{E}_{out} - \sum\dot{E}_{loss} - \sum\dot{E}_{des} = 0 \tag{5}$$

For the control volume shown in Fig. 1, the input exergy rate includes the exergy flow rate coming from the HTF and exergy rate of solar radiation which is reflected from concentrator to the receiver. The total rate of exergy input is:

$$\sum\dot{E}_{in} = \left[\dot{m}c_p\left(T_{in} - T_0 - T_0\ln\frac{T_{in}}{T_0}\right) + \frac{\dot{m}\Delta P_{in}}{\rho}\right] + \psi\eta_o I_b A_c \tag{6}$$

where, T_0 is dead state temperature, ΔP_{in} is pressure difference between inlet HTF to the receiver and ambient pressure and ψ is maximum useful work available from radiation. The amount of ψ is given by [18-22]:

$$\psi = 1 - \frac{4T_0}{T_s} + \frac{1}{3}\left(\frac{T_0}{T_s}\right)^4 \tag{7}$$

where, T_s is black body sun temperature and is considered about 5800 K [23].

The exergy output rate only include the exergy outflow rate from the HTF existing the solar receiver and is calculated from the Eq. (8)

$$\sum\dot{E}_{out} = \dot{m}c_p\left(T_{out} - T_0 - T_0\ln\frac{T_{out}}{T_0}\right) + \frac{\dot{m}\Delta P_{out}}{\rho} \tag{8}$$

For the control volume shown in Fig. 1, the rate of exergy losses is due to heat transfer losses from the solar receiver to the ambient. Therefore, the total rate of exergy losses is given by [24]:

$$\sum\dot{E}_{loss} = \dot{Q}_{loss}\left(1 - \frac{T_0}{T_w}\right) \tag{9}$$

In the solar receivers, the exergy destruction is caused by two mechanisms: exergy destruction due to HTF pressure drop through the receiver and exergy destruction due to heat transfer from high to low temperatures [24]. The rate of exergy destruction due to HTF pressure drop is as follows [25]:

$$\dot{E}_{des,\Delta p} = T_a\frac{\dot{m}\Delta p}{\rho}\frac{\ln\left(T_{out}/T_{in}\right)}{T_{out} - T_{in}} \tag{10}$$

In such systems, the exergy destruction due to heat transfer from high to low temperatures includes exergy destruction due to solar energy absorption by receiver and exergy destruction due to heat conduction from the receiver wall to the HTF. The rate of exergy destruction due to solar energy absorption is given by [26]:

$$\dot{E}_{des,abs} = \dot{m}c_P\left(T_{out} - T_{in}\right)T_0\left(\frac{1}{T_r} - \frac{1}{T_s}\right) \tag{11}$$

The rate of exergy destruction due to heat conduction from the receiver wall to the HTF is as follows [24, 27]:

$$\dot{E}_{des,cond} = \dot{m}c_p T_0\left(\ln\frac{T_{out}}{T_{in}} - \frac{T_{out} - T_{in}}{T_r}\right) \tag{12}$$

By combining Eqs. (6), (8)-(12) and general equation, Eq. (5), the exergy balance equation can be rewritten as follows:

$$\psi\eta_o I_b A_c - \dot{m}c_p\left(T_{out} - T_{in}\right)\left(1 - \frac{T_0}{T_s}\right) - \left[U_l A_{r,ap}\left(T_w - T_a\right) - \varepsilon_{eff}\sigma A_{r,\mathrm{ap}}\left(T_w^4 - T_a^4\right)\right]\left(1 - \frac{T_0}{T_w}\right) - \frac{\dot{m}\Delta p}{\rho}\left(1 - \frac{T_0}{T_{lm}}\right) = 0 \tag{13}$$

where, T_{lm} is HTF log-mean temperature difference at inlet and outlet. By combining the energy equation, Eq. (4), and exergy equation, Eq. (13), two unknown parameters; average

receiver wall temperature, T_w, and overall convective heat transfer coefficient, U_l, can be obtained..

IV. Solution Methodology

As mentioned before, to determine the average receiver wall temperature and overall convective heat transfer coefficient, two energy and exergy equations must be solved simultaneously for each experimental conditions. The properties of air, such as: density, viscosity, thermal conductivity and diffusivity are evaluated at mean temperature. A trial and error procedure is applied for solving the governing equations. First an initial value is considered for average receiver wall temperature. The air properties are evaluated at mean value of ambient and the assumed temperatures. Then, by solving energy and exergy equations (Eqs. (4) and (13)), the new value of average receiver wall temperature and the overall convective heat transfer coefficient value are obtained. In the next steps, the old receiver wall temperature is replaced by new one and the previous step is repeated again until the exact value of unknown parameters are distinct. It should be noted that all mentioned steps must be carried out for each experimental condition.

Among all the experimental data obtained in this study, the data which within their, the wind direction is parallel to concentrator plane is selected. Besides, to consider the effects of receiver geometry and on thermal analysis of system, the ensemble length scale concept from literatures [13, 14] is applied in this approach.

V. Results and Discussion

The effect of HTF mass flow rate on radiation heat loss from the receiver with different aperture diameters is shown in Fig. 2. Results indicate that, by an increasing in HTF mass flow rate the radiation heat loss decreases due to reduction in average receiver temperature.

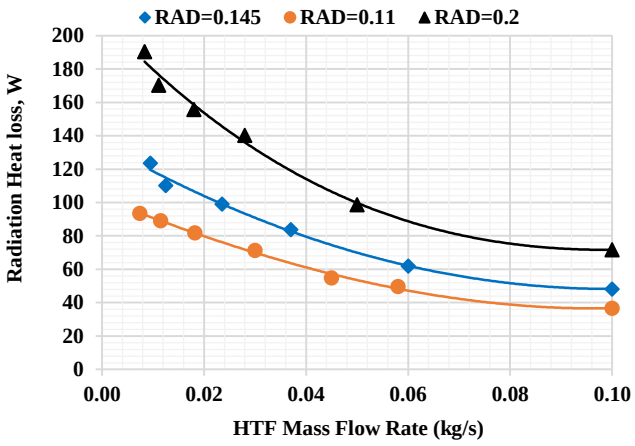


Figure 2: Effect of HTF mass flow rate on radiation heat loss for receiver with different aperture diameters, receiver emissivity=1 and I=700 W/m^2

The receiver with greater aperture diameter has greater amount of radiation loss. Explanation for this behavior is that for receiver with greater aperture the greater amount of solar radiation incident to the receiver surface and the average receiver wall temperature is greater. Because of circulating of HTF through the receiver in open path, the receiver wall temperature could not be increase so much and as a result the portion of radiation heat loss is smaller than simultaneous convective heat loss in the presence of wind. For smaller amounts of HTF mass flow rate the radiation heat loss is significant. It should be noted that the radiation heat loss cannot be neglected at low temperatures. Although, for high temperatures the radiation heat loss plays a major role.

The effect of receiver emissivity on radiation heat loss is shown in Fig. 3. By an increasing in receiver emissivity, the radiation heat loss increases. But the effect of receiver emissivity on radiation loss is so small. For low mass flow rates of HTF, where the receiver wall temperature is increased further and the radiation heat loss is more, by a decreasing in receiver emissivity about 60%, the radiation heat loss decreases only about 12.5%.

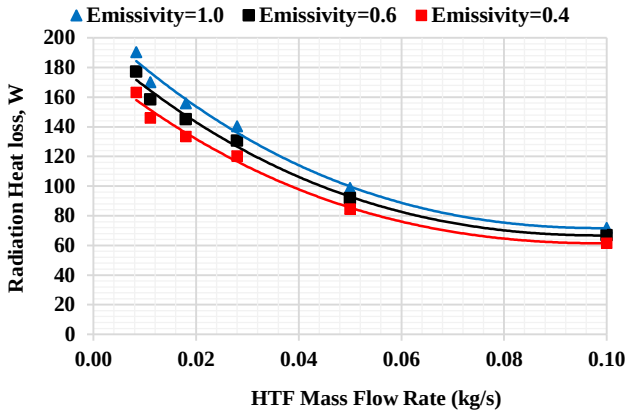


Figure 3: Effect of HTF mass flow rate on radiation heat loss for receiver with different wall emissivity for RAD=0.2 and I=700 W/m^2

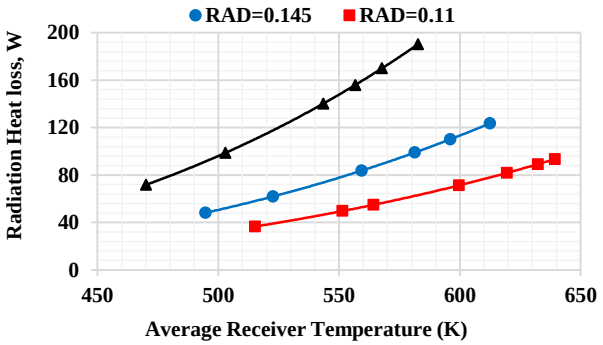


Figure 4: Effect of average receiver temperature on radiation heat loss, receiver emissivity=1 and I=1000 W/m2

The effect of average receiver wall temperature on radiation heat loss for receivers with different aperture diameters is shown in Fig. 4. Results show that the effect of receiver temperature on radiation loss for greater apertures is higher than smaller apertures. Although, for receiver with aperture diameter equal to 0.11 m, the average receiver wall temperature for small HTF mass flow rates is close to 650 K

that is higher that two other investigated cases, but, the amount of radiation heat loss is smaller due to smaller surface of heat transfer. For larger HTF mass flow rates (about 0.1 kg/s), the radiation heat loss is significantly lower for all aperture diameters and is about 70 W and the same parameter for receiver with aperture diameter equal to 0.2 m is 1.75 times of the receiver with aperture diameter equal to 0.11 m. The same ratio for small HTF mas flow rates is about 2.

The effect of average receiver temperature on convective heat loss is shown in Fig. 5. The results indicate that by increasing in receiver average temperature, due to decreasing in HTF mass flow rate, the convective heat loss increases. This figure indicates that in comparison with the variation of radiation heat loss for receiver with different receiver aperture diameters, presented in Fig. 4, the trend of convective heat loss is linear. Although, the radiation heat losses from the receiver in comparison with convection heat losses are so low, but, by decreasing the HTF mass flow rate from 0.1 to 0.0083 kg/s (through the fully opened receiver; RAD=0.2) the average receiver temperature increased from 470.08 to 582.74 K and according to Figs. 4 and 5, the convective heat loss increases only 22% while the radiation heat loss increases up to 165% through these changes.

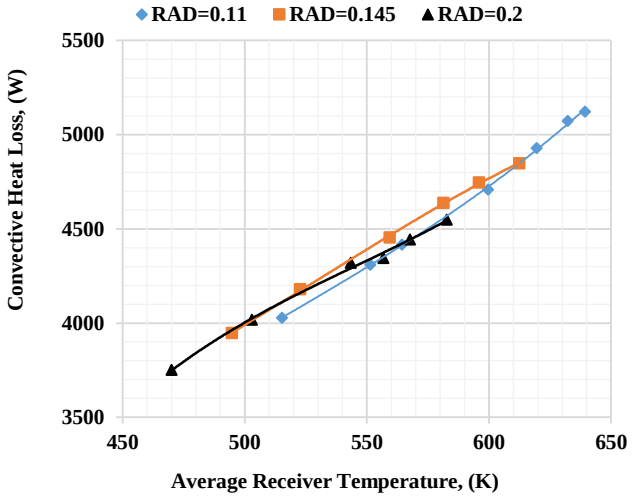


Figure 5: Effect of average receiver temperature on convection heat loss, receiver emissivity=1 and I=1000 W/m^2

The effect of HTF mass flow rate on convection heat loss for receiver with different aperture diameters is shown in Fig. 6. Results show that by an increasing in HTF mass flow rate through the receiver for all aperture diameters, the convection heat loss is reduced due to reduction in average receiver temperature. Results also indicate that for all range of HTF mass flow rates, although the fully opened receiver (RAD=0.2) provides greater heat transfer area, but the convection heat loss is less than losses of two other modes. This can be explained by using the results in Fig. 5. For solar intensity equal to 1000 W/m^2, receiver emissivity of 1 and HTF mass flow rate of 0.1 kg/s, the average receiver temperatures for RAD=0.11, 0.145 and 0.2 are 515.22, 494.74 and 470.08 K respectively. This means that the effect of receiver average temperature on heat losses is more important than the effect of receiver aperture diameters.

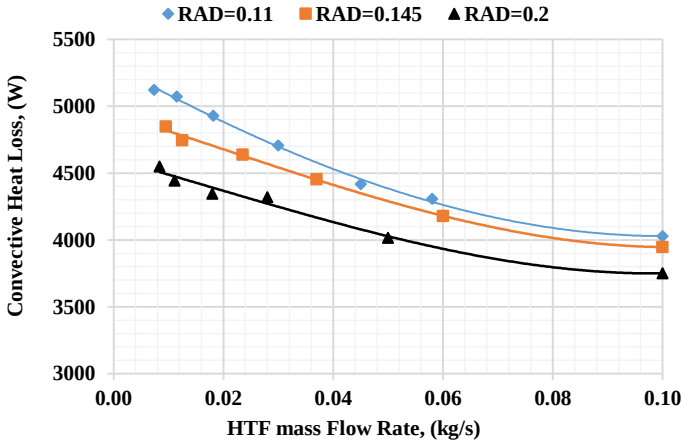


Figure 6: The effect of HTF mass flow rate on convection heat loss, receiver emissivity=1 and I=1000 W/m^2

The rate of useful gained energy or delivered energy by HTF within the receiver is defined as:

$$\dot{Q}_u = \dot{m} c_p \left(T_{out} - T_{in}\right) \tag{14}$$

The receiver thermal efficiency is defined as the ratio of the rate of delivered energy by HTF within the receiver to the rate of reflected solar radiation intensity from concentrator to the receiver and given by:

$$\eta_{t,r} = \frac{\dot{m} c_p \left(T_{out} - T_{in}\right)}{\eta_o I_b A_c} \tag{15}$$

The effect of HTF mass flow rate on receiver thermal efficiency for two aperture diameters is shown in Fig. 7. Results show that, based on the results presented in Figs. 5 and 6, the thermal efficiency is greater for receivers with greater aperture diameters, therefore the heat loss of receiver with greater apertures is less than smaller apertures. For small mass flow rates of HTF, the average receiver temperature is higher. By an increasing in HTF mass flow rate, the receiver thermal efficiency increases and reaches to a maximum value at a specified HTF mass flow rate and then decreases due to reduction in temperature difference between inlet and outlet of HTF.

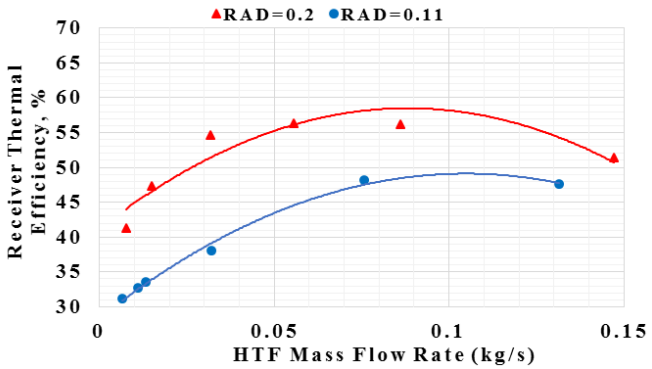


Figure 7: The effect of HTF mass flow rate on receiver thermal efficiency, receiver emissivity=1 and I=800 W/m2

The effect of wind velocity on convection heat loss is shown in

Fig. 8. Results in Fig. 8 indicate that by increasing in wind velocity the convection heat loss increases and this effect is more significant for smaller wind velocities. When the wind velocity is increased the convective heat transfer coefficient is increased and consequently the heat loss by convection increases.

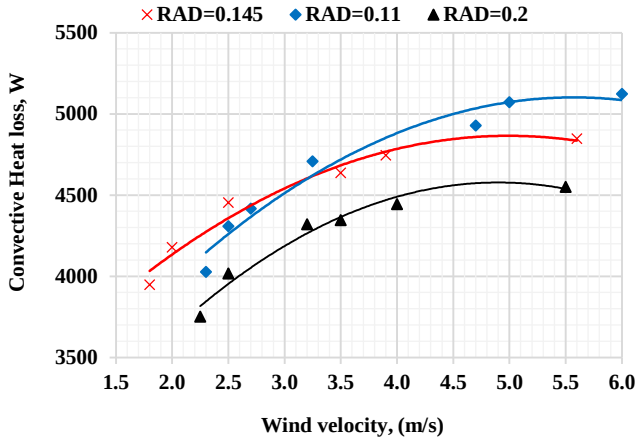


Figure 8: The effect of wind velocity parallel to concentrator plane on convection heat loss, HTF mass flow rate=0.09 kg/s, I=700 W/m2 and receiver emissivity is 1.

In parabolic dish systems, many parameters must be considered in investigation of heat losses form the receiver to the ambient. Various operational and structural parameters such as: wind velocity and direction, HTF temperature and mass flow rate, receiver geometry and inclination, affect the receiver wall temperature which is most important parameters in heat loss analysis. In this study, to consider all effects of mentioned parameters, the effect of each parameters on heat loss is deliberated separately. To take into account the effects of receiver geometry and inclination, the ensemble receiver length scale, L_s , (Ref. [13]) is applied in this analysis. The variation of *Nu* number with receiver wall temperature ratio, and *Re* is shown in Fig. 9. A generalized correlation for predicting the convection heat loss in windy condition has been developed.

A general form of Nusselt number as a function of Reynolds number and temperature ratio is considered as:

$$Nu = C\,\mathrm{Re}^{n}\left(\frac{T_w}{T_a}\right)^{m} \tag{16}$$

where, the parameters *C*, *n* and *m* are constant. By fitting the experimental data, the constant values are defined and the dimensionless convective heat transfer coefficient in term of Nusselt number is given by:

$$Nu = 1.2\,\mathrm{Re}^{0.48189}\left(\frac{T_w}{T_a}\right)^{1.9869} \tag{17}$$

$$1\times10^{3} < \mathrm{Re} < 3.7\times10^{4} \quad \& \quad 400 < T_w(K) < 950$$

where, T_w is average receiver wall temperature. Also the characteristic length in *Nu* number and *Re* number is the receiver ensemble length.

Although, Eq. (17) could not be used to predict the convective heat transfer coefficient in the cases in which that the wind direction is not parallel to concentrator plane, but it shows reasonably accurate prediction in cases with other geometry and inclination of receivers, where the wind direction is parallel to concentrator plane. The applicability of Eq. (16) is due to implementation of operational and structural parameters in correlating procedure. To verify this thread and confirmation of this claim, this correlation is applied for a conical receiver. In this investigation, the convection heat loss is estimated by using the proposed correlation for *Nu* number and the outlet HTF temperature is calculated from Eq. (1). Results show that there are a good agreement between experimental data and predicted results from thermal analysis.

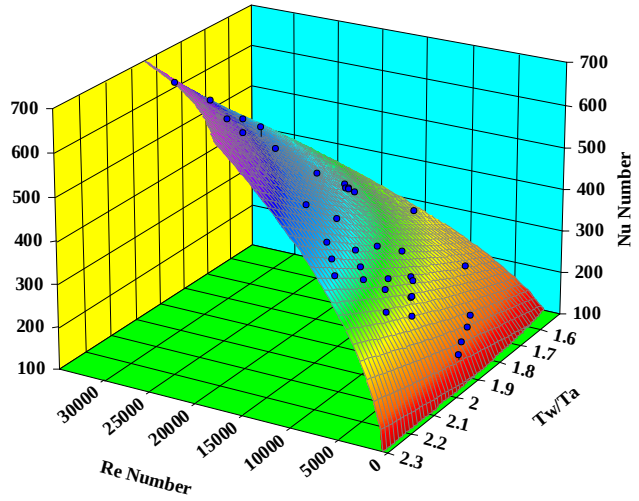


Figure 9: Variation of Nu number with receiver wall temperature ratio and Re number

Two statistical parameters are used to compare the predicted results by the proposed correlation and the experimental data. These parameters are the root mean square error, RMSE, and the coefficient of determination, R^2, which are defined as:

$$R^2 = 1 - \frac{\sum_i \left(O_i - m_i\right)^2}{\sum_i \left(O_i - \overline{O}\right)^2} \tag{18}$$

$$\overline{O} = \frac{1}{N}\sum_{i=1}^{N} O_i \tag{19}$$

$$RMSE = \sqrt{\frac{\sum_i \left(O_i - m_i\right)^2}{N}} \tag{20}$$

Here, O_i is the ith experimental data, m_i is the corresponding predicted value, *N* is the total number of data, and $\overline{O}$ is the mean value of the experimental data. The values of R^2 vary between 0 and 1, with larger numbers indicating better agreement between the predicted results and experimental data and 1 represents a perfect agreement. Also, smaller numbers of *RMSE* indicating better agreement. The R^2 and *RMSE* were calculated and results show that there is a good agreement

ISBN: 978-1-61804-253-8

between predicted results and experimental data. Numerical results related to these above mentioned statistical parameters for cylindrical receiver, are 3.79 for RMSE and the 0.98 for R^2.

In order to verify the accuracy of the proposed dimensionless convective heat transfer coefficient, the experimental data obtained from a conical receiver with two aperture diameters are used. Therefore, the predicted and experimental HTF outlet temperature versus HTF mass flow rate for a conical receiver with two aperture diameters are compared in Fig. 10. As can be seen, a good agreement between predicted and experimental values of HTF outlet temperatures is observed. It should be noted that in design of experiments for conical receiver the experimental conditions such as wind direction and range of operating parameters are considered as the same conditions for experiments of cylindrical receiver.

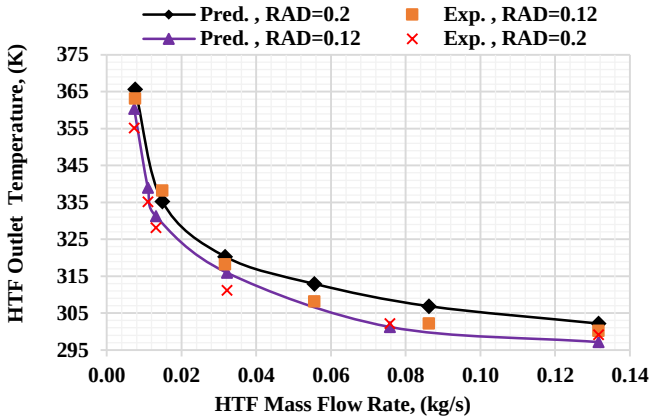


Figure 10: The effect of HTF mass flow rate on HTF outlet temperature for a conical receiver, I=1000 W/m^2

For conical receiver with RAD equal to 0.2 (fully opened condition) the values of R^2 and *RMSE* are 0.96 and 3.36 and for RAD equal to 0.11 these values are 0.95 and 3.65 respectively. The results of present study show that the proposed correlation for *Nu* number can be used for prediction of heat losses from solar cylindrical and conical cavity receivers with various geometries and inclinations, when the wind directions is parallel to concentrator plane.

VI. Conclusions

One of the most important issues in thermal performance analysis of solar receivers is heat loss estimation which plays a key role in accurate design of solar receivers and consequently system cost effectiveness. The heat is lost from the receivers by three mechanisms: conduction through receiver insulation, convection and radiation through the receiver inner surface. In most cases, the conduction heat loss is negligible due to proper insulation and the radiation heat loss can be estimated analytically. The convection heat loss is the most complicated phenomenon in thermal analysis of solar receivers and yet is a major contributor for researchers. In presence of wind, it is more complicated and important.

In this study, the heat loss from a cylindrical receiver that is employed in a parabolic dish collector was investigated. In thermal analysis, the isothermal receiver wall condition was assumed and the average receiver wall temperature is found by simultaneous energy and exergy analyses. The effects of operational parameters; HTF mass flow rate and wind speed and structural parameters, such as receiver geometry and inclination on convective heat loss are investigated. Three receiver aperture diameters for cylindrical receiver are employed.

The amount of radiative heat loss is estimate by using an analytical expression. The effect of receiver wall emissivity on radiative heat loss was studied. Results show that, compared with convective heat loss, the portion of radiative heat loss was so low and less than 10%. By increasing in receiver wall emissivity, the radiative heat loss increases, but receiver wall emissivity is not an effective parameter in heat losses. The average receiver wall temperature has significant effect on radiative heat loss, which is affected by inlet HTF mass flow rate. For small values of HTF mass flow rates, the average receiver wall temperature is higher and consequently, the radiative heat loss is greater. By an increasing in HTF mass flow rate, the average receiver wall temperature decreases and radiative heat loss is diminished for great values of HTF mass flow rates. The radiative heat loss for receiver with smaller aperture diameter was greater than two other conditions.

The convective heat loss in presence of wind was the most important issue in this work. Applying the concept of ensemble length scale, L_s, the effects of receiver geometry and inclination on convective heat loss were considered. By fitting the experimental data, a simple and empirical correlation was developed to estimate the dimensionless convective heat transfer coefficient in terms of Nu number and as a function of Re number and average receiver wall temperature. The statistical parameters including R2 and RMSE parameters are calculated to verify the accuracy of fitting procedure. The results show that the value of R2 is greater than 0.98 and the value of RMSE is smaller than 3.8 for cylindrical receiver. For validating the proposed correlation for Nu number, the data obtained through experiments with a conical receiver (with two aperture diameters) was used. The comparison of predicted values and experimental data shown that the proposed empirical correlation for Nu number can be used effectively for predicting of convection heat losses from receiver with various geometries.

References

[1] W.B. Stine, R.B. Diver, A compendium of solar dish/Stirling technology, in, DTIC Document, 1994.
[2] W.B. Stine, C. McDonald, Cavity receiver convective heat loss, in: International Solar Energy Society, Solar World Congress (1989, Kobe, Japon), 1989.
[3] T.L.I.F.P. Bergman, Fundamentals of heat and mass transfer, Wiley, Hoboken, NJ, 2011.
[4] J.P. Holman, Heat transfer, McGraw Hill Higher Education, Boston, [Mass.], 2010.
[5] A. Clausing, Convective losses from cavity solar receivers—comparisons between analytical predictions and experimental results, Journal of Solar Energy Engineering, 105 (1983) 29-33.

[6] A.M. Clausing, An analysis of convective losses from cavity solar central receivers, Solar Energy, 27 (1981) 295-300.
[7] U. Leibfried, J. Ortjohann, Convective Heat Loss from Upward and Downward-Facing Cavity Solar Receivers: Measurements and Calculations, Journal of Solar Energy Engineering, 117 (1995) 75-84.
[8] S.-Y. Wu, L. Xiao, Y. Cao, Y.-R. Li, Convection heat loss from cavity receiver in parabolic dish solar thermal power system: A review, Solar Energy, 84 (2010) 1342-1355.
[9] A. Koenig, M. Marvin, Convection heat loss sensitivity in open cavity solar receivers, Final report, DOE contract No. EG77-C-04-3985, Department of Energy, Oak Ridge, Tennessee, (1981).
[10] K. Lovegrove, T. Taumoefolau, S. Paitoonsurikarn, P. Siangsukone, G. Burgess, A. Luzzi, G. Johnston, O. Becker, W. Joe, G. Major, Paraboloidal dish solar concentrators for multi-megawatt power generation, in: ISES Solar World Conference, Goteborg, Sweden, 2003.
[11] S. Paitoonsurikarn, K. Lovegrove, On the study of convection loss from open cavity receivers in solar paraboloidal dish applications, in: Proceedings of 41st Conference of the Australia and New Zealand Solar Energy Society (ANZSES), Melbourne, Australia, 2003.
[12] S. Paitoonsurikarn, T. Taumoefolau, K. Lovegrove, Estimation of convection loss from paraboloidal dish cavity receivers, in: Proceedings of 42nd conference of the Australia and New Zealand solar energy society (ANZSES), Perth, Australia, 2004.
[13] S. Paitoonsurikarn, K. Lovegrove, A new correlation for predicting the free convection loss from solar dish concentrating receivers, in: Solar, 2006, pp. 44th.
[14] S. Paitoonsurikarn, K. Lovegrove, Effect of paraboloidal dish structure on the wind near a cavity receiver, in: Proceedings of the 44th Annual Conference of the Australian and New Zealand Solar Energy Society, Canberra, 2006.
[15] R.Y. Ma, Wind effects on convective heat loss from a cavity receiver for a parabolic concentrating solar collector, Sandia National Laboratories, 1993.
[16] V. Madadi, T. Tavakoli, A. Rahimi, First and Second Thermodynamic Law Analyses Applied to a Solar Dish Collector, Journal of Non-Equilibrium Thermodynamics, (2014).
[17] Y. Wu, L. Wen, Solar receiver performance of point focusing collector system, American Society of Mechanical Engineers, 1 (1978).
[18] W.H. PRESS, Theoretical maximum for energy from direct and diffuse sunlight, Nature 264 (1976) 734 - 735.
[19] R. Petela, Exergy of heat radiation, Journal of Heat Transfer, 86 (1964) 187-192.
[20] V. Badescu, How much work can be extracted from a radiation reservoir?, Physica A: Statistical Mechanics and its Applications, 410 (2014) 110-119.
[21] R. Feistel, Entropy flux and entropy production of stationary black-body radiation, Journal of Non-Equilibrium Thermodynamics, 36 (2011) 131-139.
[22] V. Badescu, Simple upper bound efficiencies for endoreversible conversion of thermal radiation, Journal of Non-Equilibrium Thermodynamics, 24 (1999) 196-202.
[23] A. Bejan, D. Kearney, F. Kreith, Second law analysis and synthesis of solar collector systems, Journal of Solar Energy Engineering, 103 (1981) 23-28.
[24] A. Suzuki, General theory of exergy-balance analysis and application to solar collectors, Energy, 13 (1988) 153-160.
[25] M.J. Moran, H.N. Shapiro, D.D. Boettner, M. Bailey, Fundamentals of engineering thermodynamics, (2010).
[26] A. Kahrobaian, H. Malekmohammadi, Exergy Optimization Applied to Linear Parabolic Solar Collectors, Journal of Faculty of Engineering, 42 (2008) 131-144.
[27] A. Suzuki, A fundamental equation for exergy balance on solar collectors, Journal of solar energy engineering, 110 (1988) 102-106.

Nanocrystalline $CuFeO_2$ Delafossite Thin Films Prepared on Quartz by CSP Method

Adel H.Omran Alkhayatt , S. M.Thahab , Inass Abdulah Zgair

Abstract-Transparent conductive $CuFeO_2$ thin films have been deposited on quartz substrate at 450°C using chemical spray pyrolysis technique and then annealed at 800°C for 2 h. The structural properties of $CuFeO_2$ thin films have been studied by XRD. The diffraction pattern show that $CuFeO_2$ films are polycrystalline in nature with hexagonal (Rhombohedral**)** structure. The preferred orientation along (012) plane, the grain size increases from 29.61nm to 39.566nm after annealing process. The optical properties have been investigated by UV-VIS spectrophotometer in the wavelength rang 300-800 nm, the result shows that the prepared film had transmittance at visible region about 62% and the absorption edge shifted towards long wavelength after annealing process. The direct band gap value of as deposited $CuFeO_2$ thin film was 2.66 eV and its decreases to 2.63eV for annealed film.

Key words- $CuFeO_2$, Delafossite , spray pyrolysis, thin films , TCO.

I. INTRODUCTION

Transparent conducting oxides (TCOs) are electrical conductive materials (low electrical resistance typically $<10^{-3}\Omega$.cm) [1] with a comparably low absorption of light and high optical transparency (> 80%) in the visible range of the electromagnetic spectrum [2] , with optical direct energy gaps (Eg) of > 3.1 eV [3].They are very attractive in applications such as surface acoustic wave device, varistors, gas sensors, solar cell technology, transparent conducting electrodes and in a wide variety of electronic applications[2,3,4]. Delafossite-type oxides are layered materials that consist of a variety of compositions expressed by the general formula ABO_2 A=Pt, Pd, Cu, and Ag ;B = Al, Ga, In, Sc, Cr, Fe, Co, Y, La, etc[5].The delafossite compound can exhibit both metallic and semiconductors properties depended upon the electronic structure of noble metal cation ,so elements such as Pt and Pd exhibit metallic behavior while other elements like Ag and Cu exhibit semiconductors behavior.$CuFeO_2$ is p-type of delafossite compounds that has relatively higher electrical conductivity compared with other delafossite excluding $CuCrO_2$. $CuFeO_2$ has attracted much attention as a p-type TCO used for several application such as a transparent diodes and solar cells [6]. The $Cu^{1+}Fe^{3+}O_2$ compound has a unit cell structure of hexagonal and a primitive cell structure of rhombohedral .The $CuFeO_2$ compound is an interesting thermoelectric material due to starting materials are inexpensive and non-toxic appropriate for industry [7].

II. EXPERIMENTAL DETAILS

The $CuFeO_2$ films were prepared on quartz substrates by using chemical spray pyrolysis at 450ºC and annealed at 800ºC for 2hr. Solutions of $CuCl_2{\cdot}2H_2O$ and $FeCl_3$ with 0.2 concentration were prepared by dissolved it with distilled water. The solution is sprayed onto quartz substrates with dimensions of 2.5 X 2.5 cm^2 after it ultrasonically cleaned in acetone and absolute ethanol. Each spraying period lasts for about (15 sec) following by about (3min) waiting period to avoid excessive cooling of the hot substrates duo to the spraying .The thickness of the film was found about 128 n m measured by using (Optical Thin Measurement) model (LIMF-10) by Lambda Scientific ,Ltd company. Annealing process of CuFeO2 thin film occur at 800ºC for 2hr. carried out by using Electrical Furnace (MTI Corporation) model (OTF – 1200X) by KJ group company. Crystalline structure was measured by using an X-ray diffractometer (XRD, Shimadzu -6000) using Cu Kα (λ=1,5406Å) radiation. A UV–VIS (UV-1650 Shimadzu) spectrometer was employed to determine the optical properties of the thin films in the wavelength range (380-800) nm.

III. RESULT AND DISSCUSION

A.Structural properties

The XRD patterns of the $CuFeO_2$ samples are shown inFig.1.a The characteristic peaks of all $CuFeO_2$ thin film correspond well with standard crystallographic data (JCPDS #75-2146). The main diffraction peaks of $CuFeO_2$ thin film are(101),(012),(015) with preferred

ISBN: 978-1-61804-253-8

orientation at (012) plane when the $CuFeO_2$ thin film deposited at 450ºC and After annealing crystallinity of the film was found to increase ,and the intensity of peaks improved dramatically as shown inFig.1.b . From fig.1 another phase can be found at (312)and (321)planes which belong to $CuFe_2O_4$ (JCPDS card 34 – 0425).

The lattice constant a and c for hexagonal planes of the $CuFeO_2$ thin films are calculated from XRD data using the following equation[7]:

$$\frac{1}{d^2}=\frac{4}{3}\left[\frac{h^2+hk+k^2}{a^2}\right]+\frac{l^2}{c^2} \qquad (1)$$

where (h k l) are Miller indices. The interplaner distance d_{hkl} and the full width at half-maximum (FWHM) of the diffraction peaks was calculated for as deposited and after annealing $CuFeO_2$/quartz thin films. The absorbed d_{hkl} values are in good agreement with the standard values as shown in Table1. The average grain size was calculated from the Scherrer Formula [8]:

$$D=\frac{0.9\lambda}{\beta\cos\theta} \qquad (2)$$

Where D is crystallite size (grain size), λ is the wavelength of X-ray used, β is the peak width at half maximum in radian, θ is the Bragg's diffraction angle. The grain size was found to increase with an increasing substrate temperature by annealing process(Table1). Also, the increase in average crystal size after annealing might be due to decrease in grain boundaries and the amount of defect in the films.

B. Optical properties

Fig.2 shows a typical optical transmission (T%) spectrum of$CuFeO_2$ film in the wavelength range of 380–800 nm. The average T% in the visible range for all the films is ~62%. it can be seen that the optical transmittance is influenced by annealing process and shifted towards the long wavelength (red shift).

The absorption coefficient (α) was calculated from the absorbance spectrums using the formula[9]:

$$\alpha=\frac{2.303A}{t} \qquad (3)$$

Where A and t is the absorbance and thickness of the films respectively. Absorption coefficient are increased significantly after annealing process as shown in fig.3,this can be attributed to the improvement the crystallinity of the $CuFeO_2$ thin film.

Fig.4: shows the optical band gap of $CuFeO_2$ thin films using the $(\alpha h\nu)^2$- $h\nu$ plot. A linear relationship between $(\alpha h\nu)^2$and $h\nu$ indicates that $CuFeO_2$ has a direct energy band gap. The optical absorption edge was analyzed by the following relationship[10].

$$\alpha h\nu = B(h\nu - E_g{}^{opt})^r \qquad (4)$$

where B is a constant, r value is respectively 1/2 and 2 for direct and indirect transitions. The optical direct bandgap was estimated to be 2.66 eV when deposited at 450°C by extrapolating the straight portion of the curve .When $CuFeO_2$ thin films annealed at 800 °C for 2h the optical band gap decreased to be 2.63 eV. The decreased in energy band gap after annealing can be attributed to the increasing of the grain size and the improvement the crystallinity of the film after annealing process. The reflectance R of the prepared film as deposited and annealed films was calculated using the relation :

$$R=1-A-T \qquad (5)$$

The reflectance curve shifted toward the long wavelength (low energies) for annealed film as shown in fig.5. Fig.6 shows the refractive index of $CuFeO_2$ as a function of photon energy.

The refractive index of $CuFeO_2$/quartz thin film can be calculated using Eq [9]:

$$n=\frac{1+\sqrt{R}}{1-\sqrt{R}} \qquad (6)$$

The refractive index value increased and shifted to lower energy after annealing process as shown in fig.6. The extinction coefficient is a measure of the fraction of light lost due to scattering and absorption per unit distance of the penetration medium. It can be estimated from the values of α and λ using the relation [11]|:

$$k=\frac{\alpha\lambda}{4\pi} \qquad (7)$$

The figure shows increase in extinction coefficient values gradually and shift toward long wavelengths as temperature increasing by annealing.

IV. CONCLUSIONS

The XRD studies have been showed that the CuFeO2 thin films were deposited on quartz substrates using chemical spray pyrolysis method had hexagonal (Rhombohedra) phase and the intensity of (012) direction is significantly increased with the increase of temperature by annealing process. The grain

size of the $CuFeO_2$ thin film on quartz substrate was increased after annealing process.The optical band gap for the direct optical band transition in CuFeO2thinfilms was estimated to be 2.66 eV when deposited at450º and after annealing process decreased to be2.63eV.

ACKNOWLEDGMENT

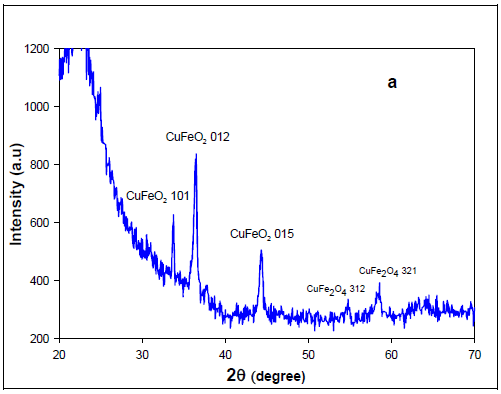
The authors would like to acknowledge the assistance offered by the Department of Physics in Faculty of Science, and NAMRU in Faculty of Engineering/University of Kufa/Iraq.

REFERENCES

1]A. Stadler ," Transparent Conducting Oxides-An Up to Date Overview ",Vol.5,PP.661-683,(2012).
[2]Z. Biju and H. Wen," Influence of substrate temperature on the structural andproperties of In-doped CdO films prepared by PLD",5 Journal of Semiconductors, Vol. 34, No. ,(2013).
[3] C. Ruttanapun," Optical and electronic properties of delafossite $CuBO_2$ p-type transparent conducting oxide", Journal of Applied Physics,Vol.114, PP.113108 (2013).
[4] A. V. Chadwick, A. N. Blacklocks, A. Delafossite-type $CuInO_2$ Thin Films", Journal of Physics: Conference Series,Vol.249,PP.012045, (2010).
[5] M. Yasukawa , K. Ikeuchi, T. Kono, K. Ueda and H.Hosono " Thermoelectric properties of delafossite-type layered oxides $AgIn_{1-x}Sn_xO_2$",Journal of Applied Physics98, 013706 (2005).
6] M. M. Moharam, M. M. Rashad, E. M. Elsayed and R. M. Abou Shahba ,"A acile novel synthesis of delafossite $CuFeO_2$ powders,"Journal of Materials Science: Materials in Electronics Volume 25, Issue 4, PP .1798-1803 ,(2014).
[7]C.Rudradawong ,A.Wichainchai ,A. Sakulkalavek,Y.Hongaromkid and C.Ruttanapun ," Method of High Active Preparation and Electrical Properties of $CuFeO_2$ delafossite – type", Journal of Advanced Material Research ,Vol.979,PP. 302-306,(2014).
[8] S.R. Elliott, "Physics of Amorphous materials", Long man Group limited (1983).
materials", Long man Group limited (1983).
[9] J.I.Pankove, " Optical Process in Semiconductors ", Dover Publishing, Inc., New York. (1971).
[10] .Tauce, "Amorphous & Liquid Semiconductors", Plenum, London & New York,(1974).
[11] S.M.Sze and Kwok K. Ng,"Physics of Semiconductor Devices Third Edition, John Wiley & Sons, Inc. ,(2007).

Table1: Obtained results from the XRD for $CuFeO_2$ thin films.

NO		2θ degree	d(A°) measure	d(A°) Standard	(hkl)	Grain Size n m	Average Grain Size n m	a (A°)	C (A°)	FWHM (degree)
1	$CuFeO_2$ deposited at 450°C	34.5587	2.5936	2.5982	(101)	G1=53.78	29.61	3.01	16.74	0.15
		35.9947	2.4932	2.5133	(012)	G2=14.29				0.59
		43.9303	2.0596	2.0870	(015)	G3=20.76				0.41
2	$CuFeO_2$ deposited at 450°C and annealed at 800°C	34.4124	2.6041	2.5982	(101)	G1=65.99	39.566	3.01	16.68	0.13
		35.9964	2.4932	2.5133	(012)	G2=28.588				0.29
		43.9783	2.0574	2.0870	(015)	G3=24.12				0.13



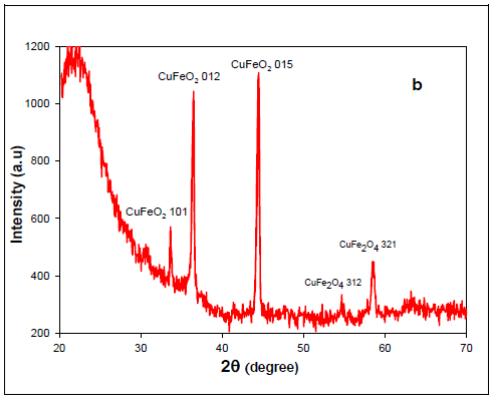


Fig.1: XRD pattern for $CuFeO_2$ thin films prepared using chemical spray pyrolysis (a: T_S = 450ºC (b: T_a = 800 ºC.

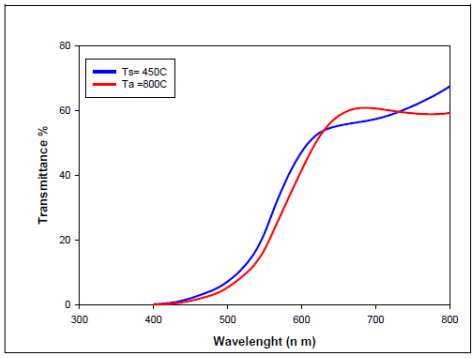


Fig.2 : Transmittance spectra of $CuFeO_2$/quartz thin films deposed at T_s= 450 ^oC and annealed at T_a=800^o C.

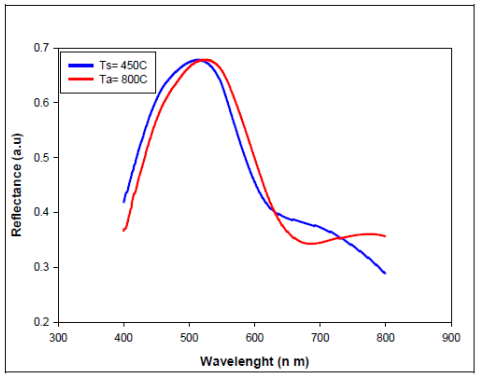


Fig.5: Reflectance spectra of $CuFeO_2$/quartz thin film deposed at T_s=450 ^oC and annealed at T_a=800 ^oC.

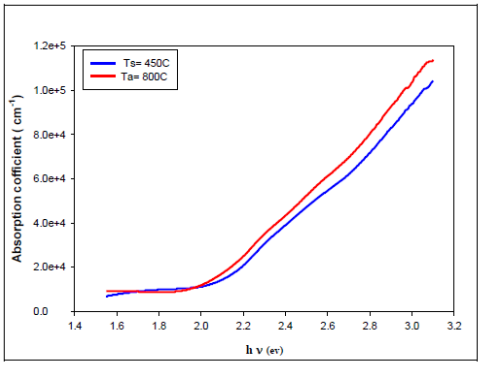


Fig.3 Absorption coefficient of $CuFeO_2$/quartz thin films deposed at T_s= 450 ^oC and annealed at T_a = 800 ^{o} C.

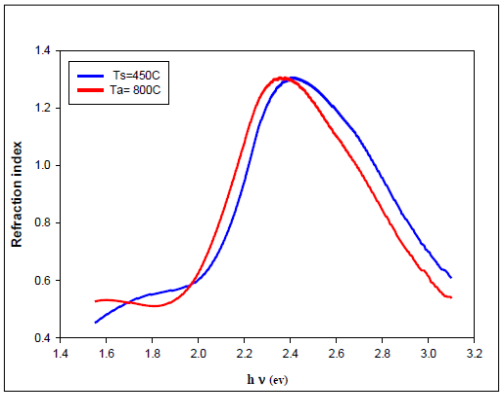


Fig.6:Refractive index of $CuFeO_2$/quartz thin film deposed at T_s=450^{o}C and annealed at T_a=800^{o}C.

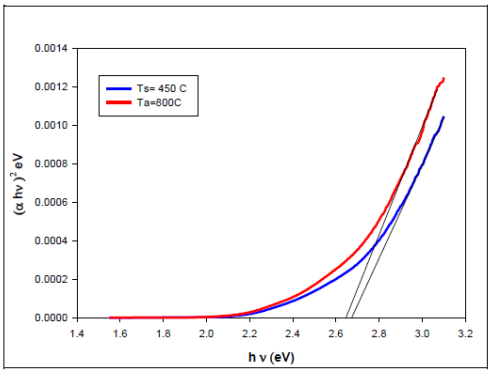


Fig.4: Direct energy band gap of $CuFeO_2$/quartz thin film deposed at T_s=450C and annealed at T_a= 800C.

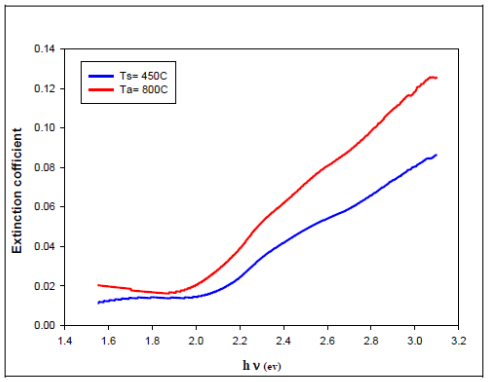


Fig.7 Extinction coefficient of $CuFeO_2$/ quartz thin film deposed at T_s=450^{o}Cand annealed at T_a=800^{o}C

Relative level of magnetizing granular matrix samples varying in length: calculating dependences

Sandulyak A. A., Sandulyak A. V., Ershova V. A.

***Abstract* –** We have studied and realized phenomenological and physical approaches to obtaining formulae for calculating relative level of magnetizing cylindrical filter-matrix samples of various lengths. We have shown a considerable influence of the samples relative dimensions on this level becoming quite low with small values of the relative size, which is characteristic for filter-type separators used in production. There are corresponding calculating formula (for narrowed, close to real, ranges of the relative dimension and adequate ranges of demagnetizing factor), the formulae being applicable for solving a wide array of direct and inverse problems of magnetic separation.

***Keywords* –** demagnetizing factor, form factor, granular filter-matrix, relative dimension, relative level of magnetization.

I. Introduction

MATRIX magnetic separators of filter type [1 – 5] are used to remove ferro-impurities from various technological media in order to conform to the standard quality indicators. A specific operating element, viz. a magnetized matrix performs the targeted capture of ferroimpurities in such apparatuses; to be more exact, the loading of granules in the working volume of the separator serves this function. However, there is little attention paid to the issue of evaluating the extent to which potential capabilities of a granular matrix used as a magnetic are employed. Thus, there are errors in identifying the operating regimes, inconsistencies in factual and expected results of such apparatuses, incomplete and not always objective (including comparative) estimates of the results obtained.

For instance, when choosing particular values of the volume and the corresponding matrix sizes, people often overlook the following important condition. Magnetic properties of a matrix-body, e.g. such fundamental magnetic (and therefore, magnetic-sorption) parameters as a mean magnetic induction B_b and/or magnetic permeability μ_b, except for magnetic properties of the material (matter) of the granules and their concentration in the volume given, greatly depend on the form of this body, this holds true for a solid magnetic as well. In particular, for a cylindrical body these parameters vary with its length L, to be more specific, they vary with the relation of this length to the cylinder diameter D, i.e. a relative dimension of L/D as a peculiar form factor. It is this circumstance, ignored as a rule, which causes the so-called demagnetizing factor N worsening magnetic properties of a matrix in comparison with potential values of induction B and/or magnetic permeability μ, these properties are also inherent for a quite long or classic toroidal sample of the same matrix.

The issue of obtaining such vitally necessary information on the demagnetizing factor of non-homogeneous (granular, grain and other) magnets as distinct from solid ones have not been sufficiently elaborated until recently. Actually, a systematic study of the matter was initiated in [6]. Thus, as applied to the sample bodies of these magnets (a loading of ball-bearing balls) of a cylindrical form of various lengths, i.e. with different relative dimension L/D, field dependences of B_b induction have been obtained experimentally. They have really demonstrated quite strong influence of parameter L/D on a relative level of induction $B_b/B=\Lambda$ and permeability $\mu_b/\mu=\Lambda$, especially at relatively low (let us note, intrinsic to filter-separators) values of L/D (Fig. 1*a*), when Λ value can amount to just Λ =0.5-0.6 (50-60%) and less. At that, equation $\mu_b/\mu= B_b/B=\Lambda$ holds true on the ground of $B_b= \mu_0\mu_b H$ and $B=\mu_0\mu H$, where μ_0 is a magnetic constant, H is the intensity of a magnetizing field; while the relations μ_b/μ and B_b/B per se can be conveniently named by a relative level of magnetization Λ as an indicator of fractional usage of potential magnetic properties of a magnet.

In connection with the objectively arising need for a routine accounting of the matrix relative dimension L/D real role in the level of its magnetization Λ in a particular case, it becomes necessary to obtain the corresponding calculating dependencies. With their help basing on a given and/or assumed geometry of the filter-matrix, we could provide fair and accurate operational information about factual and anticipated working efficiency (in terms of filter-matrix magnetization) of the existing and newly created separators. We can also solve a number of problems related to magnetic separation, direct and inverse ones.

This work was supported by Russian Federation Ministry of Education and Science No. 9.1189.2014.

A. A. Sandulyak, PhD in Technical Sciences, is with Moscow State University of Instrument Engineering and Computer Science (MGUPI), Russian Federation (corresponding author to provide phone: 007-903-723-52-44; e-mail: anna.sandulyak@mail.ru).

A. V. Sandulyak, professor, is with Moscow State University of Instrument Engineering and Computer Science (MGUPI), Russian Federation (e-mail: a.sandulyak@mail.ru).

V. A. Ershova, PhD in Technical Sciences, is with Moscow State University of Civil Engineering (MGSU), Russian Federation, (e-mail: v.ershova@mail.ru).

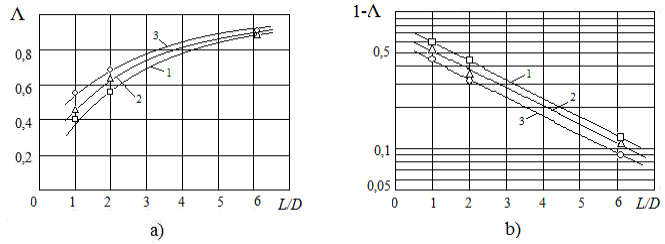


Fig.1. A relative level of magnetization of filter-matrix samples (*a*), and the shortage of this level (*b*) varying with their relative dimension by data in [6] (points), 1–*H*=30 (kA/m) (μ=8.5), 2–*H*=60 (kA/m) (μ=7.4), 3–*H*=80 (kA/m) (μ=6.9). The lines are the calculation according to (6) and (8).

II. Research findings on elaboration and implementation of various approaches to obtaining formule for calculating relative level of filter-matrix magnetization

Elaboration of a phenomenological approach (without legalising a demagnetization factor as a quantitative parameter).

In research [7], the field dependences obtained in [6] for the mean induction in the granular matrix samples were subjected to a pertinent processing with the use of semi-logarithmic coordinates convenient for the purpose. As a result, without resorting to obtaining the quantitative data on the demagnetization factor (demagnetizing coefficient) *N* for a characteristic range of magnetizing field intensity *H*=30–80 (kA/m), we have managed to find a phenomenological expression allowing us to define a relative level of filter-matrix magnetization Λ depending on its relative dimension *L/D* [7]:

$$\Lambda = 1 - A_\mu \, exp\left(-0{,}35\, L/D\right), \qquad (1)$$

Herewith in this equation parameter A_μ was assumed as being a constant averaged value $A_\mu \cong 0.75$.

It is easy to show (and this is well seen by the layering of experimental dependencies in Fig.1) that in order to define formula (1) the averaging of obtained dependencies of the magnetization level shortage performed in [7], i.e. (1-Λ), on *L/D* (the averaging with the accuracy to $A_\mu \cong 0.75$) may serve only as the first approximation. In practice, there is a functional, even though weak connection between parameters A_μ and *H*, at least for the aforementioned range of *H* it may be expressed by a power function (*H* in kA/m):

$$A_\mu \cong 1/H^{0,07}. \qquad (2)$$

In this connexion, expression (1) should be written as follows:

$$\Lambda = 1 - \frac{exp\left(-0{,}35\, L/D\right)}{H^{0,07}}. \qquad (3)$$

Together with that, the upgraded calculations, even though they conform to the found in [7] exponential nature of connection between the shortage of relative magnetization level (1-Λ) and the relative dimension *L/D*, but in a slightly corrected form (Fig. 1*b)* may be presented as:

$$\Lambda = A_\mu \, exp\left(-0{,}32\, L/D\right), \qquad (4)$$

according to the received updated data of A_μ directly read from Fig. 1*b* with *L/D* →0, indicate a different exponential character (Fig. 2*a*) of parameter A_μ on *H* dependence (*H* – in kA/m):

$$A_\mu = exp\left(-0{,}0062H\right). \qquad (5)$$

Then, from (4) and (5), alternatively to (3), there follows one more expression for a relative level of magnetization:

$$\Lambda = 1 - exp\left(-0{,}0062H - 0{,}32\, L/D\right). \qquad (6)$$

Provided we operate not with the parameter of magnetising field intensity *H*, but employ a parameter of mean magnetic permeability of the matter of this quasi-solid medium μ which depends on *H*, the values are given in Fig.1 legend, the same A_μ values (Fig.2*a*)

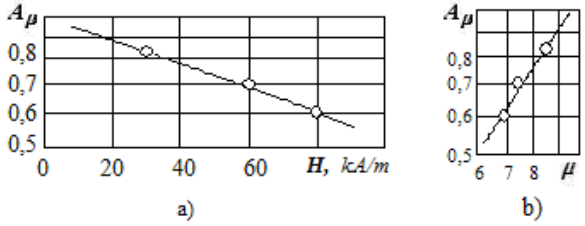


Fig. 2. A_μ parameter data in (4) represented in semi-logarithmic (*a*) and logarithmic (*b*) coordinates to illustrate respective exponential and power dependencies of this parameter on the magnetising field intensity and magnetic permeability of the 'matter' of the quasi-solid filter-matrix.

can be expressed by μ (Fig.2*b*), and we can obtain a functional connection close to a power one:

$$A_\mu = 0{,}033\mu^{1,5}. \qquad (7)$$

In this case alongside with expression (6) we can put down a related expression:

$$\Lambda = 1 - 0{,}033\mu^{1,5}\, exp\left(-0{,}32\, L/D\right). \qquad (8)$$

Expressions (6) and (8) which in the frameworks of a phenomenological model represent the solutions of an important problem of identifying a relative level of magnetizing "shortened" and "short" filter-matrixes of magnetic separators correspond to the experimental data depicted in Fig. 1*a*, which, basically, served the foundation for obtaining these solutions.

Implementation of a physical approach (with legalising the demagnetization factor as a quantitative parameter).

The same fundamental connection between the demagnetization factor (a demagnetizing coefficient) *N* and the magnetic permeability of the matter μ and the body μ_b [8-12] can be applied to a granular (formally – a quasi-solid) magnet, as to a solid one:

$$N = \frac{1}{\mu_b - 1} - \frac{1}{\mu - 1}. \qquad (9)$$

Thereat, it is easy to get a calculating formula from this connection to define a relative level of filter-matrix magnetization; just to remind you – it is a formula for a relative level of magnetic permeability and respective magnetic induction $\mu_b/\mu = B_b/B = \Lambda$:

$$\Lambda = \frac{1}{\mu}\left[\frac{\mu - 1}{(\mu - 1)N + 1} + 1\right], \qquad (10)$$

needless to say, the formula can be obtained for a given sample of a filter-matrix with known values of magnetic permeability of its quasi-solid matter μ and a demagnetization factor N.

Note that in contrast to the afore-considered phenomenological approach where the demagnetizing factor N of the sample-body was a kind of a shadow parameter, exhibiting itself latently as an externally observed phenomenon of sample-body magnetization level reduction, here N appears as a quantitative parameter. Its influence on Λ is evaluated quite definitely (Fig.3), namely, in the range μ= 5-10 covering a characteristic for granular matrixes range of μ [6] dependencies of Λ on N considerably decrease with N increase (Fig.3).

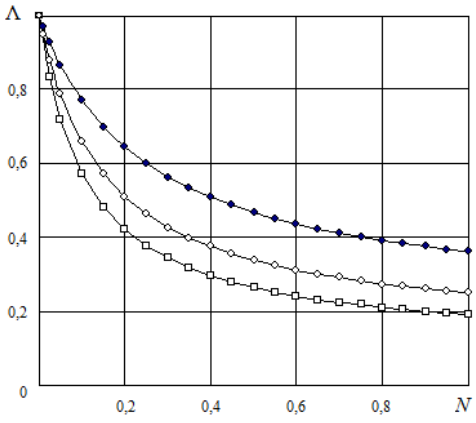


Fig. 3. Relative level of magnetic body magnetizing depending on its demagnetizing factor 1 – μ=5; 2 – μ=7.5; 3 – μ=10.

As to the data necessary for calculating N values by formula (10) as applied to the specific samples of granular medium, they are amply given in paper [13] for the samples of a cylindrical granular magnet (balls loading). Thus, following the definition (9) and basing on the aforementioned experimental data in [6] and in paper [13] there have been performed calculations for specific values of parameter N for various values of L/D of such samples. Moreover, appropriate processing of these data manifested, that the dependence of the demagnetization factor N on the sample relative dimension L/D has an exponential view, but with a quite peculiar argument, viz. a radical of the relative dimension [13], i.e.

$$N = exp\left(-k_N\sqrt{L/D}\right), \qquad (11)$$

with the defined value of coefficient $k_N \cong 1.5$ in (11) true for the studied quasi-solid (granular) samples in the range μ= 6.9-8.5 [6] characteristic for the magnetizing separator matrixes.

Consequently, taking into account relation (11), expression (10) written in the expanded view as:

$$\Lambda = \frac{1}{\mu}\left[\frac{\mu - 1}{(\mu - 1)exp\left(-1{,}5\sqrt{L/D}\right) + 1} + 1\right], \qquad (12)$$

becomes a basic calculating formula which fully characterises the influence of immediate 'initiator' of parameter N, a relative dimension of filter-matrix L/D on the relative level of its magnetization Λ.

The dependencies obtained with expression (12) shown in Fig.4 for the investigated range H=30-80 (кА/м) (μ=8.5–6.9) are in accord with dependences

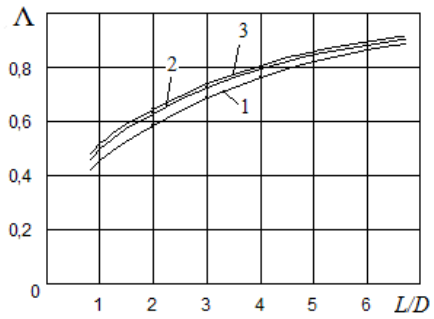


Fig. 4. Calculation data obtained according to (12), the data of relative magnetization level of filter-matrix samples in terms of their relative size, 1–μ=8.5, 2–μ=7.4, 3– μ=6.9.

shown in Fig. 1*a* and with the expressions (6) and (8) ensuing from them. It signifies a good agreement of the respective approaches considered, phenomenological and physical ones.

Formalising the results of the physical approach for practical (comparatively narrow) range of the demagnetizing factor.

Expression (10) following from the classical relation (9) and facilitating the elaboration of the calculating formula (12) is naturally true for the entire range of possible varying of the demagnetization factor: N=0-1. Alongside with that, for filter-matrixes of industrial and experimental magnetic separators relative dimension L/D values of which are mainly in the order of unity and more, the real range of N values is relatively narrow: N < 0.2-0.25 [13]. That is why it seems reasonable to try to obtain its particular, simplified phenomenological variants for the narrowed range of N, basing on expression (10).

To do this, we should refer to the earlier studied family of dependencies Λ on N (Fig.3) constructed according to expression (10), but as the argument it is rational to use not N but $N^{0,7}$ (Fig.5), similarly to the method used in [13] to obtain expression (11) with argument $(L/D)^{0,5}$.

In these coordinates and in the range of μ=5-10, which is still of great interest for us, the aforementioned dependencies are linearised quite well, as is seen in Fig. 5*a*, but only in the range of $N^{0,7}$< 0.25-0.4 (Fig.5*a*, dashed lines) which corresponds to somewhat reduced values of N, namely N< 0.14-0.27. It means that in this case the calculating formula has the following view:

ISBN: 978-1-61804-253-8

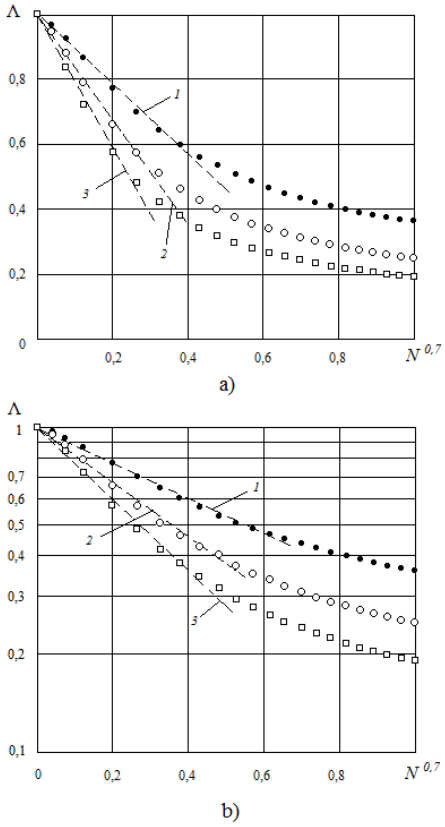


Fig. 5. On the possibility of fractional linearization of data in Fig. 3 in common (*a)* and semi-logarithmic (*b*) coordinates, the argument is a power function of the demagnetizing factor.

$$\Lambda = 1 - k_\mu N^{0,7}, \tag{13}$$

with that parameter k_μ depending on the magnetic permeability μ of the quasi-solid magnetic 'matter' (Fig. 5*a*, dashed lines) complies to the linear connection with μ (Fig. 6, curve 1):

$$k_\mu = 0{,}21\mu, \tag{14}$$

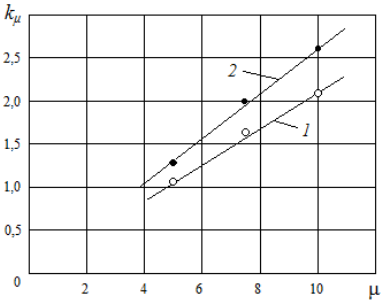


Fig. 6. Illustration of linear dependence of k_μ coefficient introduced in formulae (13) and (17), lines 1 and 2 respectively, on magnetic permeability of the quasi-solid filter-matrix 'matter'.

which allows us, basing on (13), to put down the following version of the expression for the relative level of filter-matrix magnetization (for the ranges N<0.14-0.27 and μ=5-10):

$$\Lambda = 1 - 0{,}21\mu N^{0,7}. \tag{15}$$

Subsequently, inserting here expression (11) for N we may obtain, alternatively to expression (12), the following calculating formula for the stated ranges of N and μ*:*

$$\Lambda = 1 - \frac{0{,}21\mu}{exp\sqrt{L/D}}. \tag{16}$$

From the point of view of considerable expansion of N values range, i.e. expansion to the values of $N^{0,7}$< 0.45-0.65, or up to a quite acceptable range of N< 0.32-0.54, it is more preferable to employ the option of semi-logarithmic coordinates used for representing in them the same data (shown in Fig. 5*a*). The fact of fractional (for the mentioned range of N) linearization of these data used in semi-logarithmic coordinates (Fig. 5*b*, dashed lines) indicates the possibility of using the calculating formula of an exponential type:

$$\Lambda = exp\left(-k_\mu N^{0,7}\right), \tag{17}$$

whereas parameter k_μ, depending on the magnetic permeability μ of the quasi-solid magnetic 'matter' (Fig. 5*a*, dashed lines) also complies to the linear but slightly different than before connection with μ (Fig. 6, curve 2):

$$k_\mu = 0{,}26\mu. \tag{18}$$

After substituting (18) in (17) there follows one more version of the expression for the relative level of filter-matrix magnetization, here it is for the stated above expanded range of N< 0.32-0.54 and the previous range of μ=5-10:

$$\Lambda = exp\left(-0{,}26\mu N^{0,7}\right). \tag{19}$$

And substitution of (11) into (19) yields, parallel with expressions (12) and (16), the following calculating (formal) formula for the above stated ranges of N and μ:

$$\Lambda = exp\left(-\frac{0{,}26\mu}{exp\sqrt{L/D}}\right). \tag{20}$$

The obtained formulae (16) and (20) may be used for practical calculations when solving direct and inverse problems connected with identifying actual level of magnetization Λ of a specific (by L/D) filter-matrix of the magnetic separator, and which is nonetheless important, the tasks related to defining the necessary value of the filter-matrix relative dimension L/D:

$$\frac{L}{D} = \left(ln\frac{0{,}21\mu}{1-\Lambda}\right)^2, \quad \frac{L}{D} = \left[ln\left(-\frac{0{,}26\mu}{ln\,\Lambda}\right)\right]^2, \tag{21}$$

proceeding from the pre-set level of its magnetization Λ.

For example, when creating an apparatus (experimental, pilot production, or industrial one) there is a decision taken to allow the reduction of the filter-matrix magnetization level by no more than say 20%, i.e. the acceptable value of Λ for this case is Λ=0.8. Then the calculation by formula (21), let's say for the case of μ =7.5, manifests the necessity to procure obligatory relative dimension of the working member of this apparatus, i.e. filter-matrix, with L/D=4.1-4.7 dimension. By the way, similar result (L/D=4.2) is received when performing calculations with the control formula (12) after turning L/D parameter from an argument into a function.

ISBN: 978-1-61804-253-8

III. Conclusion

Judging by the data of real relative level of magnetizing granular matrix samples with various values of such parameter as their relative dimension *L/D*, we have to acknowledge the form factor role and dependant on it demagnetizing factor of samples (as quasi-solid magnets) to be essential and compulsory to be taken into account. Factual data on this parameter of magnetic separators filter-matrixes geometry bespeak of apparent under-employment of their magnetic properties. For instance for a filter-matrix the length of which is commensurable with its diameter, a relative level of its magnetization is no less than 50 – 60%, whereas the corresponding indicator of under-usage of the same filter-matrix potential capabilities amounts to 40 – 50%. Or, in other words, 40 – 50% of the separator magnetising system capabilities is ineffective, dead-weight and actually wasted.

Thus, the calculating formulae necessary for developers and operators of the filter-type magnetic separators and connecting the form factor of the filter-matrix (a relative dimension) with its magnetization are in higher demand, especially in the context of the increasing demand for magnetic separators. The considered in this paper approaches to obtaining such formulae and the solutions found to some extent compensate for the existing gap in the field, which allows expanding the circle of solvable direct and inverse problems of magnetic separation.

With the help of the received designs, it becomes possible to get objective data on the filter-matrix capacity (almost always it is a 'short' one) as a magnet, and thus, on the corresponding sorption screen for ferroparticles, both in newly developed and already operating magnetic separators. This being said, it is also possible to provide concept estimation of the reasons for frequent inconsistencies in the magnetic separation results, including those in relation to the announced anticipated values stipulated in passport, technical and advertising materials and the like. Furthermore, we can facilitate the tasks connected with designing separators of such a type, e.g. by obtaining the data on acceptable values of relative dimensions of the filter-matrix as a primary working body for the filter magnetic separator.

References

[1] A. Newns, R. D. Pascoe, "Influence of path length and slurry velocity on the removal of iron from kaolin using a high gradient magnetic separator," *Minerals Engineering*, 2002, Vol. 15, pp. 465-467.

[2] J. G. Rayner, T. J. Napier-Munn, "A mathematical model of concentrate solids content for the wet drum magnetic separator," *International Journal of Mineral Processing*, 2003, Vol. 70, pp. 53-65.

[3] V. Zezulka, P. Straka, P. Mucha, "A magnetic filter with permanent magnets on the basis of rare earth," *Journal of Magnetism and Magnetic Materials*, 2004, Vol. 268, pp. 219-226.

4. A. A. Sandulyak, A. V. Sandulyak. "Prospects of employing magnetic filter-separators for purifying ceramic suspensions," *Glass and Ceramics*, 2006, No.11, pp. 34-37.
5. D. Norrgran. "Magnetic filtration: producing fine high-purity feedstocks" *Filtration and Separation*, 2008, Vol. 45 (6), pp. 15-17.
6. A. V. Sandulyak. "Purification of liquids in a magnetic field," *Higher School Publishing House at Lvov University*, 1984, 167 p.
7. A. V. Sandulyak, C. Saccani, A. A. Sheipak. "Basic criteria and fundamentals of constructing magnetic filters," *Heavy engineering*, 2000, No.3, pp. 31-38.
8. D.-X. Chen, J. A. Brug, R. B. Goldfarb. "Demagnetizing factors for cylinders," *IEEE Transactions on Magnetics*, Vol. 27, No. 4, 1991, pp. 3601-3619.
9. R. Goleman R. "Macroscopic model of particles' capture by the elliptic cross-section collector in magnetic separator," *Journal of Magnetism and Magnetic Materials*, 2004, Vol. 272-276, pp. 2348-2349.
10. K. Smistrup, O. Hansen, H. Bruus, M. Hansen. "Magnetic separation in microfluidic systems using microfabricated electromagnets – experiments and simulations," *Journal of Magnetism and Magnetic Materials*, 2005, Vol. 293, pp. 597-604.
11. D.-X. Chen, E. Pardo, A. Sanchez. "Fluxmetric and magnetometric demagnetizing factors for cylinders," *Journal of Magnetism and Magnetic Materials*, 2006, Vol. 306, pp. 135-146.
12. K. Nandy, S. Chaudhuri, R. Ganguly, I. K. Puri. "Analytical model for the magnetophoretic capture of magnetic microspheres in microfluidic devices," *Journal of Magnetism and Magnetic Materials*, 2008, Vol. 320, pp. 1398-1405.
13. A. A. Sandulyak, V. A. Ershova, D. V. Ershov, A.V. Sandulyak. "On the properties of short granular magnets with unordered granule chains: a field between the granules," *Solid State Physics*, 2010, Vol. 52, Issue 10, pp. 1967-1974.

Application of Highly Stereoselective Co-Catalytic Direct Aldol Reaction on Water for the Concise Synthesis of D-*lyxo*-Phytosphingosine

Moniruzzaman Mridha,[a,b] Guangning Ma,[a] Carlos Palo-Nieto[a] and Armando Cordova*[a,c]

***Abstract*—** An efficient, stereoselective and concise synthetic route to D-*lyxo*-phytosphingosine has been achieved. The key feature in this strategy was the highly stereo- and *syn*-selective direct aldol reaction catalyzed by combining acyclic amino acid and hydrogen bond donor catalysts in the presence of water. The synthesis was developed starting from commercially available TBS protected dihydroxyacetone as donor and pentadecanal as acceptor in the presence of water to get the D-*lyxo*-phytosphingosine in 29 % overall yield (4 steps).

***Keywords*—** Aldol reaction, Asymmetric synthesis, Organocatalysis, Phytosphingosine , Reductive amination.

I. Introduction

THE aldol reaction is one of the most powerful methods to form C-C bonds in synthetic organic chemistry.[1] In nature, Type I and Type II aldolase enzymes catalyze the aldol reaction with excellent stereoselectivity via a catalytic enamine and a Zn-enolate mechanism, respectively.[2] Inspired by nature, chemists started to design small organic molecule catalysts for the asymmetric aldol reaction. In this context, the development of catalytic methods for the enantioselective aldol reaction is a highly active research area.[3] Nowadays, organocatalysis is playing a very significant role in asymmetric synthesis.[3d, 3e] In this context, proline and its derivatives are excellent organic catalysts for asymmetric aldol reaction.[4] In 1970s, Hajos and Parrish disclosed the first proline-catalyzed aldol reaction.[4a] This reaction was used several times for total synthesis applications and in the pharmaceutical industry. Then almost 30 years later List, Barbas and Lerner disclosed the use of proline as a catalyst for the intermolecular asymmetric aldol reaction.[4d] Inspired by these works, a wide range of small organic molecules including proline and their chiral derivatives have been used as catalysts for the asymmetric aldol reaction.[3] Later, our group disclosed that acyclic amino acids and small non-proline derived peptides could also catalyze the intermolecular aldol reaction with high enantioselectivity.[5] We found that water accelerated these amino acid and peptide catalyzed reactions. In 2005, Gellman and coworkers reported the co catalytic effect using hydrogen bond donor to form C-C bond via Michael addition reaction.[6a] Later on, Shan et al. reported the co-catalytic aldol reaction catalyzed by proline using (*S*)-BINOL as the additive.[6b] Proline catalyzed aldol reactions are highly *anti*-selective. However, in 2007, Lu reported highly *syn*-selective aldol recations catalyzed by an *O*-TBS protected threonine derivative.[7a] After that, Barbas showed that O-*t*Bu protected threonine was a highly syn-selective organocatalyst. Recently, Li and his group demonstrated a large-scale asymmetric direct aldol reaction employing threonine derivatives as recoverable organocatalyst.[7c] Recently, our group reported a highly efficient and enantioselective asymmetric aldol reaction using hydrogen bonding donor as a co-catalyst along with acyclic amino acid derivatives in organic solvent.[7d]

Phytospingosines are one of the major components of the structural backbone of spingolipids containing 1-amino-2,3-diol moiety, which play a significant role in several physiological processes.[8a] Phytosphingosine was first isolated from mushrooms in 1911[8b] and now it has been investigated that it is widely distributed in the microorganisms, plants and even several mammalian tissues for instance, hair, kidney,[8c] skin,[8d] liver,[8e] and in blood plasma. Phytosphingosine itself is a bioactive lipid for example D-*ribo* phytospingosine is a heat stress signal in yeast cells[8f] and induces apoptosis in cancer cells.[8g] Phytosphingosine is also backbone of KRN7000; a α-glactosylphytoceramide enhanced killer activities and strongly inhibited tumor metastasis in mice[8h] (Fig. 1).

M.M. Author is with Department of Natural Sciences, Mid Sweden University, SE-851 70 Sundsvall, Sweden; and Department of Chemistry – BMC, Uppsala University, SE-751 23 Uppsala, Sweden.

G.M. and C.P.N. Authors are with Department of Natural Sciences, Mid Sweden University, SE-851 70 Sundsvall, Sweden. A.C. is with with Department of Natural Sciences, Mid Sweden University, SE-851 70 Sundsvall, Sweden; and Berzelii Center EXSELENT, The Arrhenius Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden. (Phone: +46 707415178; fax: +46-8-154908; e-mail: armando.cordova@ miun.se).

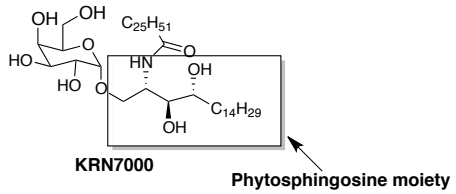


Fig. 1: α-glactosylphytoceramide KRN7000

Four stereoisomers of phytosphingosine are found in nature (Fig. 2).

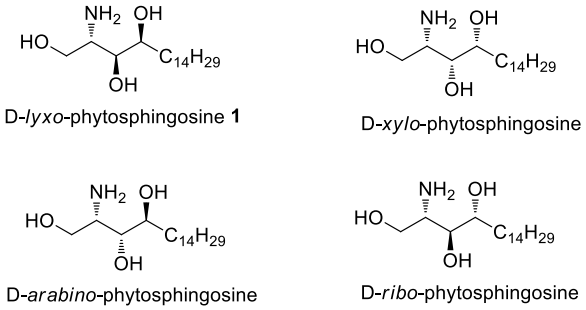
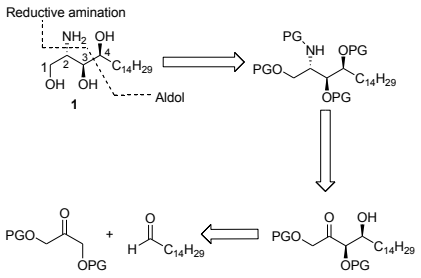


Fig. 2: Naturally occurring phytosphingosine stereoisomers.

Due to its promising biological activity, the construction of aminotriol moiety with desired stereochemistry has become a synthetic challenge. Although several approaches of the target phytosphingosine **1** and other stereoisomers have already been reported, most of the methods suffer ether from expensive starting material, several synthetic steps, extensive protecting group manipulations, low yield, stereo or regeoselectivity.[9]-[12] Therefore, a concise, stereoselective and high yield synthetic route of target phytosphingosine stereosisomers are still demandable. Recently, Enders reported the elegant synthesis of *arabino*- and *ribo*-phytosphingosine stereosisomers by the proline-catalyzed *anti*-selective aldol reaction.[13] Jørgensen has also disclosed an elegant catalytic one-pot approach to these compounds.[14]

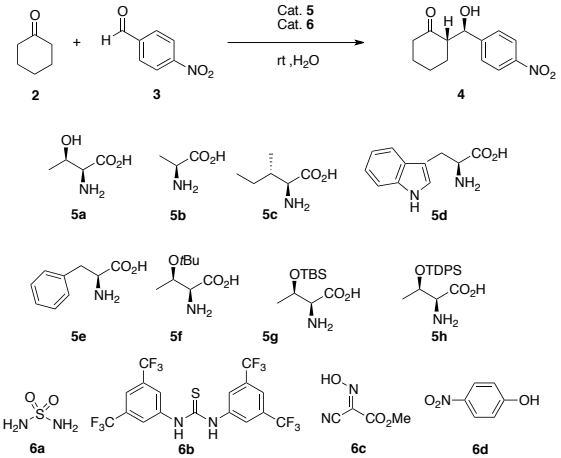
In this paper, we wish to report a highly enantioselective direct aldol reaction on water by combining hydrogen bond-donating and acyclic amino acid catalysts. This transformation using protected dihydroxyacetone as the donor was next applied as the key step for the total synthesis of D-*lyxo*-phytosphingosine **1** (structural backbone of spingolipids) starting from inexpensive and commercially available starting materials.

Synthetic plan: On the basis of the retrosynthetic design we constructed C_3-C_4 by a potential *syn*-selective aldol reaction. Then the bulky protected group was introduced at the hydroxyl group. The bulkiness of the protected group on the secondary alcohol was responsible for controlling the diastereoselectivity in reductive amination step. This process efficiently adjusts the stereoselectivity and simultaneous deprotection would generate the desired D-*lyxo* phytosphingosine **1**. (Scheme. 1)



Scheme 1: Retrosynthetic design to synthesis of D-*lyxo*-[2*S*, 3*S*, 4*S*]-phytosphingosine **1**. PG = protective group.

The direct co-catalytic intermolecular aldol reaction was carried out between cyclohexanone **2** and 4-nitrobenzaldehyde **3** in the presence of different chiral acyclic amino acid catalysts and hydrogen-bond donating catalysts on water. Water was chosen as a solvent because we wanted to avoid the use of toxic organic solvent as well as water has proven accelerate and improve the efficiency of direct catalytic aldol reactions when hydrophobic substrates have been employed.[5c,15-17] Moreover, water is available, safe and environment friendly. We found that hydrophobic amino acids **5** could catalyze the reaction in water (Scheme 2).

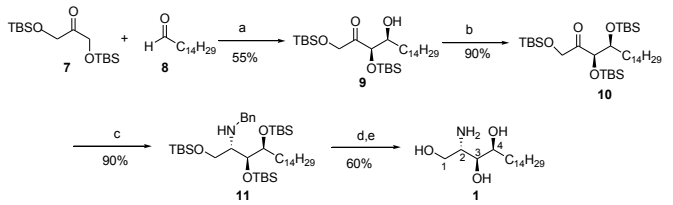


Scheme 2. Catalyst asymmetric aldol reaction

The best stereoselectivity was observed when the protected threonine derivative **5g** was used as the co-catalyst. In addition, the significant rate acceleration was observed when the reactions were performed in the presence of hydrogen-bond donating catalysts **6c** and **6d**, respectively. For example, the employment of **5g** and gave corresponding aldol product **2** in 93% yield with dr 18:1 and ee 98% after 4h. Adding **6c** as the co-catalyst gave **2** in 92% yield with 19:1 dr (*anti*:*syn*) and 98% ee after 2h. In fact, the latter was the best example of the screening. Thus, the combination of amino acid **5g** with hydrogen bond donor **6c** was chosen as the best co-catalyst system for the stereoselective aldol reaction in the presence of water. With these results in hand, we embarked on the total synthesis of D-*lyxo* phytosphingosine **1** (Scheme 3). The synthesis of was initiated by employing the co-catalytic *syn*-selective aldol reaction starting from commercially available

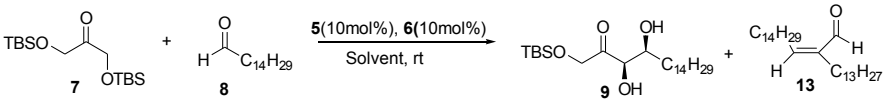
ISBN: 978-1-61804-253-8

TBS protected dihydroxyacetone **7** as the donor and pentadecanal **8** as the acceptor using amino acid **5g** and **6c** as the catalysts in the presence of water.



Scheme 3. *Reagents and conditions:* a) *O*-(TBS)-L-threonine (**5g**), Oxime (**6c**), H_2O, rt, 24h. b) TBSOTf, 2,6-lutidine, DCM, -15˚C to rt. c) $BnNH_2$, AcOH, 4Å MS, $NaBH_3CN$, DCM, -15˚C to rt, 24h. d) Pd/C, H_2 (1 atm), MeOH, rt, 24h e) TBAF, THF, rt, 24h.

The co-catalytic *syn*-selective aldol recation gave the corresponding *syn* aldol product **9** in good yield (55%) and with 19:1 (*syn/anti*) after 24 h. On the basis of our previously reported article on co-catalytic direct aldol reaction,[7d] we also tried the combination of catalyst **5g** with hydrogen bond donor **6b** in toluene. However, the result was not satisfactory as we were able to isolate the desired aldol product **9** in 43% yield with a lower *syn/anti* ratio (12:1) after 48 h. (Table I). Thus, a significant beneficial effect for the stereoselectivity as well efficiency of the aldol recation was observed. The lower yield may be explained by the fact that the linear aldehydes also undergo self-aldol condensation, which indirectly competed with the desired cross-aldol reaction. The formation of an aldol product with high *syn*-stereochemistry is consistent with related *O*-protected threonine derivatives catalyzed *syn*-aldol reactions.[7]



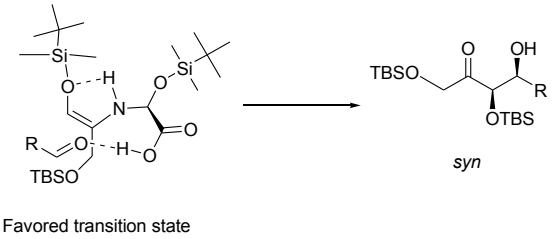
Entry	5	6	solvent	Time (h)	yield(%)	dr (*syn:anti*)[a]
1	**5g**	**6c**	H_2O	24	55[b]	19:1
2	**5g**	**6b**	Toluene	48	43[c]	12:1

a) the dr was determind by ^{1}H-NMR of the crude mixture. b) Self aldol condensation **13** has been isolated 19% .

c) Self aldol condensation **13** has been isolated 22% .

Table I: Stereoselective *syn* aldol reaction

Previously, it was reported that protected primary amino acid catalyzed asymmetric aldol reactions are either *anti* or *syn*-selective depending on the ketone donor.[7a-7d] When the donor is cyclic the aldol product has an *anti*-configuration. However, when the donor is acyclic the aldol product has a *syn*-stereochemistry. We therefore propose that the primary amino acid-catalyzed the *syn*-selective aldol reaction via the six-membered transition state involving an carboxylic acid-catalyzed enamine mechanism (Scheme.4).[18]



Scheme 4: the proposed transition state for the syn-configuration.

We next proceed with synthesis of 1. Thus, protection of the alcohol functionality of aldol **9**[19] by a bulky group should make it possible to regulate the reductive amination step. The conversion of the alcohol group to the corresponding TBS-ether **10** was accomplished with TBSOTf and 2,6-lutidine in dichloromethane at -15˚C to afford **10** in 90% yield. The TBS-ether was converted into the corresponding amine derivative **11** by a highly diastereocontrolled reductive amination using sodium cyanoborohydride as the reducing agent and benzyl amine as the amino donor together with acetic acid in dichloromethane at -15 °C. The reaction proceeded smoothly to furnish the protected amine **11** in 90% yield with excellent diastereoselectivity (*anti*:*syn* 26:1, determined by ^{1}H NMR analysis). Thus, the neucleophilic addition of the hydride across the imine occurs selectively from the *Re*-face to generate the *anti*–cofiguration (C_2-C_3). The presence of a bulky *R* group hinders and disfavors the approach of the hydride species to the imine from the *Si* face, which would then lead to the *syn*-configuration (C_2-C_3). Finally, the removal of the benzyl group by hydrogenation and the subsequent deprotection gave the desired product D-*lyxo* phytosphingonine **1** in 64 % yield.

II. Conclusion

The highly stereoselective co-catalytic direct aldol reaction between cyclohexanone **2** and 4-nitrobenzaldehyde **3** by the combination of amino acid and a hydrogen bonding donor in the presence of water has been explored. The corresponding aldol product **4** was formed in high yield (up to 93%) with high diastereoselectivity (*anti*) and enantioselectivity (up to 99%). This methodology was applied in a stereoselective *syn* -aldol reaction to generate the corresponding aldol product **9** with high *syn*-configuration (19:1 dr, C_3-C_4) Subsequent protection of compound **9** by a sterically demanding group gave **10** and subsequent diastereoselective reductive amination gave the corresponding compound **11** containing the desired configuration with high dr (26:1). Finally complete deprotection afforded the desired D-*lyxo*-phytosphingosine **1** with an overall yield of 29% in 4 steps. So far this is the most efficient and commercially viable

route to synthesis of D-*lyxo*-phytosphingosine **1 when** comparing with other methods.

ACKNOWLEDGMENT

Financial support was provided by Mid Sweden University and the Swedish National Research Council (V R).

REFERENCES

[1] a) B.M. Trost, I. Flemming, C.-H. Heathcock (Eds.), *Comprehensive Organic Synthesis,* Vol 2, Pergamon, Oxford, **1991**; b) R. Mahrwald (Ed.), *Modern Aldol Reactions*, vols. 1 and 2, Wiley-VCH, Weinheim, Germany, **2004**; For applications in total synthesis, see: c) D.A. Evans, D.M. Fitch, T. E. Smith, V.J. Cee, *J. Am.Chem. Soc.* **2000**, *122*, 10033-10046; d) K.C. Nicolaou, D. Vourloumis, N. Winssinger, P.S. Baran, *Angew. Chem. Int. Ed*. **2000**, *39*, 44-122. e) T. Mukaiyama, *Angew. Chem. Int. Ed*. **2004**, *43*, 5590-5614; f) B. Schetter, R. Mahrwald, *Angew. Chem. Int. Ed.* **2006**, *45*, 7506-7525.

[2] a) R. Beisswenger, G. Snatzke, J. Thiem, M. R. Kula, *Tetrahedron Lett.* **1991,** *32*, 3159–3162. b) T. D. Machajewski, C.-H. Wong, *Angew. Chem. Int. Ed.* **2000**, *39*, 1352-1375; c) R. N. Patel (Ed.), *Stereoselective Biocatalysis*, Marcel Dekker, New York, **2000**.

[3] For selected reviews for enantioselective aldol reaction see: a) P. I. Dalko, L. Moisan, *Angew. Chem. Int. Ed*. **2001**, *40*, 3726-3748. b) B. List, *Tetrahedron* **2002**, *58*, 5573-5590; c) P. I. Dalko, L. Moisan, *Angew. Chem. Int. Ed*. **2004**, *43*, 5138-5175; d) A. Berkessel, H. GrÖger (Eds.), *Asymmetric Organocatalysis*, Wiley-VCH, Weinheim, **2005**, Germany; e) P. Dalko (Ed.), *Asymmetric Organocatalysis*, Wiley-VCH, Weinheim, **2007**, Germany.

[4] For the early examples of proline-catalyzed direct aldol reactions, see: (a) Z. G. Hajos and D. R. Parrish, *J. Org. Chem.* **1974**, *39*, 1615-1621; (b) U. Eder, R. Sauer, R. Wiechert, *Angew. Chem. Int. Ed.* **1971**, *10*, 496-497; (c) C. Pidathala, L. Hoang, N. Vignola, B. List, *Angew. Chem.,Int. Ed.* **2003**, *42*, 2785; (d) B. List, R. A. Lerner, C. F. Barbas III, *J. Am. Chem. Soc.* **2000**, *122*, 2395-2396.

[5] a) A. Córdova, W. Zou, I. Ibrahem, E. Reyes, M. Engqvist, W.-W. Liao, *Chem. Commun*. **2005**, *28*, 3586-3588; b) W. You, I. Ibrahem, P. Dziedzic, H. Suden, A. Córdova, *Chem. Commun*. **2005**. 4946-4948; c) P. Dziedzic, W. Zou, J. Hafren, A. Córdova, *Org. Biomol. Chem.* **2006**. *4*, 38-40. d) A. Córdova, W. Zou, P. Dziedzic, I. Ibrahem, E. Reyes, Y. Xu, *Chem. Eur. J*. **2006**, *12*, 5383–5397.

[6] For pioneering work of using hydrogen bond donor to improve the stereoselectivity of catalytic aldol reactions by enamine activation, see: a) T. J. Peelen, Y. Chi, S. H. Gellman, *J. Am. Chem. Soc.* **2005**, *127*, 11598-11599; b) Y. Zhou, Z. Shan, *J. Org. Chem.* **2006**, *71*, 9510-9512; c) Y. Chi, S. H. Gellman, *Org. Lett.* **2005**, *7*, 4253-4256; For a review, see: d) X. Companyo, M. Vicanio, R. Rios, *Mini-Rev. Org. Chem.* **2010**, *7*, 1-9.

[7] For primary amino acid catalyzed *syn*-selective direct aldol reaction see: a) X. Wu, Z. Jiang, H. M. Shen, Y. Lu, *Adv. Synth. Catal.* **2007**, *349*, 812-816; b) S. S. V. Ramasastry, H. Zhang, F. Tanaka, C. F. Barbas III , *J. Am. Chem. Soc.* **2007**, *129*, 288-289; c) C. Wu, X. Fu, S. Li, *Eur. J. Org. Chem*. **2011**, *2011*, 1291–1299; d) G. Ma, A. Bartoszwicz, I. Ibrahem, A. Córdova, *Adv. Synth. Catal.* **2011**, *353*, 3114-3122; For dihydroxyacetone-based *syn*-selective direct aldol reactions see: e) S.S.V. Ramasastry, K. Albertshofer, N. Utsumi, C. F. Barbas III, *Org. Lett*. **2008**, *10*, 1621-1624; f) S.S.V. Ramasastry, K. Albertshofer, N. Utsumi, F. Tanaka, C.F. Barbas III, *Angew. Chem. Int. Ed.* **2007**, *46*, 5572-5575; g) N. Utsumi, M. Imai, F. Tanaka, S.S.V. Ramasastry, C.F. Barbas III, *Org. Lett.* **2007**, *9*, 3445-3448; (h) M. Markert, M. Mulzer, B. Schetter, R. Mahrwald, *J. Am. Chem. Soc.* **2007**, *129*, 7258-7259; (i) M. Markert, R. Mahrwald, *Eur. J. Chem.* **2008**, *14*, 40-48.

[8] a) A.R. Howell, A.J. Ndakala, *Curr. Org. Chem*. **2002**, *6*, 365-391; b) J. Zellner, *Monatsh. Chem.* **1911**, *32*, 133; c) K.A. Karlsson, B.E. Samuelsson, G.O. Steen, *Acta Chem. Scand.* **1968**, *22*, 1361-1364; d) P.W. Wertz, M.C. Miethke, S.A. Long, J.S. Stauss, D.T. Downing, *J. Invest. Dermatol*. **1985**, *84*, 410-412; e) D.E. Vance, C.C. Sweeley, *J. Lipid Res*. **1967**, *8*, 621-630; f) R.C. Dickson, R.L. Lester, *Biochim. Biophys. Acta.* **2003**, *1583*, 13-25; g) M. T. Park, J.A. Kang, J. A. Choi, C. M. Kang, T. H. Kim, S. Bae, S.Kang, S.Kim, W. I. Choi, K. C. Cho, H. Y. Chung, Y. S. Lee, S. Lee, *J. Clin. Cancer Res.* **2003**, *9*, 878-885; h) E. Kobayashi, K. Motoki, T. Uchida, H. Fukushima, Y. Koezuka, *Oncol. Res*. **1995**, *7*, 529-534.

[9] For recent synthesis of D-*lyxo*-phytosphingosine, see: a) G. Righi, S. Ciambrone, C. D. Achille, A. Leonelli, C. Bonini, *Tetrahedron*, **2006**, *62*, 11821-11826; b) S. Kim, N. Lee, S. Lee, T. Lee, Y. M. Lee, *J. Org. Chem*. **2008**, *73*, 1379-1385; c) Y. M. Lee, D. J. Baek, S. Lee, D. Lim, S. Kim, *J. Org. Chem.* **2011**, *76*, 408-416; d) Y. Mu, J.Y. Kim, X. Jin, S. H. Park, J.E. Joo, W.H. Ham, *Synthesis*, **2012**, *44*, 2340-2346; e). G.S. Rao, B.V. Rao, *Tetrahedron Lett.* **2011**, *52*, 6076-6079; f). E. Abraham, E.A. Brock, J.I. Candela-Lena, S.G. Davies, M. Georgiou, R.L. Nicholson, J.H. Perkins, P.M. Roberts, A.J. Russell, E.M. Sanchez-Fernandez, P.M. Scott, A.D. Smith, J.E. Thomson, *Org.Biomol.Chem.***2008**, *6*, 4668-4669.

[10] For other methods towards the synthesis of *lyxo*-phytospingosine, see: a) X. Lu, H.S. Byun, R. Bittman, *J. Org. Chem.* **2004**, *69*, 5433-5438; b) T. Tsujimoto, Y. Itoa, *Tetrahedron Lett.* **2007**, *48*, 5513-5516; c) I. Kumar, C.V. Rode, *Tetrahedron: Asymmetry* **2007**, *18*, 1975–1980; d) J. Park, J.H. Lee, Q. Li, K. Diaz, Y.T. Chang, S. K. Chung, *Bioorg. Chem.* **2008**, *36*, 220-228; e) A. Dubey, P. Kumar, *Tetrahedron Lett.* **2009,** *50*, 3425–3427.

[11] For a review of synthesis of phytosphingosines and their analogues, see: a) J.A. Morales-Serna, J. Llaveria, Y. Diaz, M.I. Matheu, S. Castillón, *Curr. Org. Chem.* **2010**, *14*, 2483–2521 and references cited therein; b) A.R. Howell, A. Ndakala, *J. Curr. Org. Chem.* **2002,** *6*, 365–391 and references cited therein.

[12] Selected publications for syntheses of phytosphingosines, see: a) A.J. Ndakala, M. Hashemzadeh, R.C. So, A.R. Howell, *Org. Lett.* **2002**, *4*, 1719-1722; b) S. Raghavan, A. Rajender, *J. Org. Chem.* **2003**, *68*, 7094-7097; c) R.C. So, R. Ndonye, D.P. Izmirian, S.K. Richardson, R.L. Guerrera, A.R. Howell, *J. Org. Chem.* **2004**, *69*, 3233-3235; d) P. Bhaket, K. Morris, C.S. Stauffer, A. Datta, *Org. Lett.* **2005**, *7*, 875-876; e) J. Liu, Y. Du, X. Dong, S. Meng, J. Xiao, L. Cheng, *Carbohydr. Res.* **2006**, *341*, 2653-2657; f) L.V.R. Reddy, P.V. Reddy, A.K. Shaw, *Tetrahedron: Asymmetry* **2007**, *18*, 542-546; g) Y. Ichikawa, K. Matsunaga, T. Masuda, H. Kotsuki, K. Nakano, *Tetrahedron* **2008**, *64*, 11313-11318; h) G.S. Reddipalli, M. Venkataiah, M.K. Mishra, N.W. Fadnavis, *Tetrahedron: Asymmetry* **2009**, *20*, 1802-1805; i) P. Kumar, A. Dubey, V.G. Puranik, *Org. Biomol. Chem.* **2010**, *8*, 5074-5086; j) A. S. Eleuterio, L. Quintero, F.S. Piscil, *J. Org. Chem.* **2011**, *76*, 5466-5471.

[13] D. Enders, J. Palecek, C. Grondahl, *Chem. Commun*. **2006**, 655-657.

[14] Jiang et. al. *Angew. Chem. Int. Ed*. **2009**, *48*, 6844-6848.

[15] (a) S. Naryan, J. Muldoon, M.G. Finn, V.V. Fokin, H.C. Kolb, K.B. Sharpless *Angew*. *Chem., Int*. *Ed*. **2005**, *44* 2005, 3275. (b) Y. Jung, R. A. Marcus *J*. *Am*. *Chem*. *Soc*. **2007**, *129*, 5492.

[16] For reviews discussing factors required to make a reaction in water environmentally benign, see: (a) A. Chanda, V. V. Fokin, *Chem. Rev*. **2009**, *109*, 725; (b) D. G. Blackmond, A. Armstrong, V. Coombe, A. Wells, *Angew. Chem., Int. Ed*. **2007**, *46*, 3798.

[17] P. Dziedzic, A. Bartoszewitcz, A. Córdova, *Tetrahedron Lett.* **2009**, *50*, 7242.

[18] A. Bassan, W. Zou, E. Reyes, F. Himo, A. Córdova, *Angew. Chem. Int. Ed*. **2005**, *44*, 7028.

[19] **9**: $[\alpha]^{20}_D$ = +3.1° (*c* = 1.0, $CHCl_3$). IR(neat): 2924 (s), 2853 (s), 1734 (C=O) ,1463(m) ,1388 (w), 1361(w), 1252(m), 1096(m), 1005(w), 938(w), 834(s), 776(s), 733(w), 674(w). ^{1}H NMR (500 MHz, $CDCl_3$): δ 4.50 (d, *J* = 18.4 Hz, 1H), 4.45 (d, *J* = 18.4 Hz, 1 H), 4.30 (d, *J* = 2.8 Hz, 1H), 3.77 (m, 1H), 2.16 (d, *J* = 9.8 Hz, 1H), 1.47 (m, 2H), 1.35-1.20 (m, 24H), 0.94 (s, 9H), 0.90 (s, 9H), 0.86 (t, *J*=7.0 Hz, 3H), 0.09 (s, 9H), 0.06 (s, 3H). ^{13}C NMR(125.8 MHz, $CDCl_3$): δ 210.5, 79.2, 73.2, 68.6, 34.1, 32.1, 29.85, 29.84, 29.82, 29.79, 29.71, 29.6, 29.5, 26.0, 25.96, 25.89, 22.85, 18.6, 18.3, 14.2, -4.6, -4.9, -5.2, -5.3.

[20] D-*lyxo*-Phytosphingosine **1**: white solid m.p. 106-107°C (lit[9b] 104-105°C, lit[9d] 104.5-105.5°C) $[\alpha]_D^{20}$= -9.5 (*c* = 0.96, pyridine) (lit[9d] $[\alpha]_D^{20}$ = -6.4 (*c* = 1.0, pyridine)); IR(neat): 3346 (br s) , 2922 (s) , 2852 (m) , 2360 (m), 2341 (w), 1733 (w), 1586 (w), 1464 (m), 1102 (s). ^{1}H NMR (500 MHz, pyridine-d_5): δ 4.33 (m, 2H), 4.24 (m, 1H), 4.10 (m, 1H), 3.73 (m, 1H), 2.02 (m, 1H), 1.88 (m, 1H), 1.72 (m, 1H), 1.56 (m, 1H), 1.45-1.15 (m, 22H), 0.85 (t, *J* = 6.3 Hz, 3H). ^{13}C NMR (125.28 MHz, pyridine-d_5): δ 74.4, 72.0, 64.2, 56.7, 34.6, 32.1, 30.2, 30.1, 30.0, 29.9, 29.6, 26.7, 22.9, 14.3; HRMS (ESI^+) $[M+H]^+$ Calcd for $C_{18}H_{40}NO_3$: 318.3003, found: 318.3005.

Fabrication and Characterization of SOFC components by Spray Pyrolysis method and Conventional methods

G. Tsimekas, E. Papastergiades and N.E. Kiratzis

***Abstract*—**The technique of spray pyrolysis (SP) was utilized to fabricate thin films of electrodes and electrolytes suitable for operation in Solid Oxide Fuel Cells. Anodic electrodes were chosen for their suitability as suitable electrocatalysts for the direct oxidation of hydrocarbons in these high temperature fuel cells. Attention was given to the optimization of this technique in terms of process parameters and their effect on film morphology and crystal structure. Electrochemical characterization showed encouraging results for SP fabricated anodic films on dense electrolyte pellets.

***Keywords*—**SOFCs, Spray Pyrolysis, ceramic films, YSZ,cermets,LSM.

I. Introduction

SOLID oxide fuel cells exhibit the advantages of high efficiency and low pollutants emissions. One of the greatest obstacles to wide commercialization is their high fabrication and materials cost [1]. Those cells operate at high temperatures (500-1000°C) using as typical electrolyte zirconia stabilized with yttria $Zr_{0.92}Y_{0.08}O_{1.96}$ (YSZ) which is a good oxygen ion (O^{2-}) conductor at high temperatures. Alternatively, electrolyte of ceria (CeO_2) stabilized with gadolinia (Gd_2O_3) (CGO) can be used which has higher ionic conductivity in comparison to YSZ at temperatures 500-750°C.

Reliability and stability of those cells can be optimized by reducing the operating temperature and control the microstructure of components which will assure high cell performance at low temperatures. The fabrication method of the cell influences those two parameters. The Spray Pyrolysis (SP) technique consists of a spraying solution of ionic salts on a suitable substrate at certain temperature for the production of oxides in film form [2]. The advantage of the method lies in the accurate control of stoichiometry at droplet level and its simplicity. In addition, it allows for the direct observation of the impact of the process parameters on the final product. The method has been used for the production of solid oxide fuel cell (SOFC) components with planar geometry [3-6]. Also it can be applied for the fabrication of electrolytic films with low thickness resulting in lower ohmic losses and for the fabrication of all the cell components in situ.

In the present work results are presented of the application of this method for the production of anodic film cermets such as Cu-CeO_2, Co-CeO_2 and Cu-$La_{0.75}Sr_{0.25}Cr_{0.5}Mn_{0.5}O_{3-\delta}$ (Cu-LSCM) and cathodic electrodes of $La_{0.75}Sr_{0.25}MnO_3$ (LSM) on YSZ substrates as well as electrolytic YSZ films on LSM substrates. These anodes are suitable anodic electrodes for the direct oxidation of hydrocarbons in SOFC [7]. In addition composite structures of Cu-LSCM/YSZ were fabricated on LSM substrates. The produced films were characterized for morphology by X-ray diffraction and scanning electron microscopy. Anodic films by spray pyrolysis showed optimized electrochemical performance in comparison with those fabricated by conventional methods.

II. Experimental

A. Experimental apparatus

The experimental apparatus of Fig. 1 was used for the spray pyrolysis deposition experiments. The system consists of a syringe pump (KD-Scientific, model KDS-100) which controls the solution flow rate during spraying, an air compressor (vol.50L) with manometer and a flow regulator valve which controls air pressure and air flow rate. The two fluids (solution – air) are mixed in a spray nozzle (JUJ3, Spraying Systems Co.) which produces a spray with droplets diameter in the range of 10 – 100 μm. The substrate was placed on a ceramic hot plate the temperature of which is measured by a NiCr-Ni

This research is implemented through the Operational Program "Education and Lifelong Learning" and is co-financed by the European Union (European Social Fund) and Greek national funds. Research Program: ARCHIMEDES III-Investing in knowledge society, Managing Authority: Ministry of Education & Religious Affairs, (NSRF: 2007-2013)

G. Tsimekas is a researcher at the Technological Education Institute of West Macedonia (TEIWM), Kozani, Greece and PhD candidate at the Chemistry Department, University of St. Andrews, Scotland with Professor JTS Irvine (e-mail: gtsimekas@ gmail.com).

E. Papastergiades is with the Department of Food Technology and Nutrition at the Alexander Technological Education Institute of Thessaloniki, Thessaloniki Greece (e-mail: efspa@food.teithe.gr).

N. E. Kiratzis is professor at the Department of Environmental and Pollution Control Engineering of TEIWM, Kozani, Greece (phone: +30-24610-68143; fax: +30-24610-39682; e-mail: kiratzis@teiwm.gr).

thermocouple (K-GREISINGER electronic, Germany). The temperature data during spraying were recorded by a suitable interface software to a PC. The distance between the spray nozzle and the substrate was kept constant at 20 cm. After the spraying was completed, the produced films sintered in a high temperature tube furnace (Thermoconcept) at 700 – 1000°C for 4-5 hours with heating-cooling rates 2°C /min.

B. Materials and Solutions

Anodic ceramic-metal electrodes Cu-CeO_2, Cu-LSCM and Co-CeO_2 had the same composition i.e. 3:7 (v/v metal: ceramic) according to percolation theory [8]. Nitrate salts were used for the preparation of precursor solutions at suitable ion composition ratios at concentrations 0.025M-0.2M except for Mn for which $C_4H_6MnO_4 \cdot 4H_2O$ was used. Distilled water was used as solvent (<2.5 μS/cm). Anodic films were deposited on either electrolytic pellets of Scandia stabilized zirconia (ScSZ) provided with screen printed LSM cathode on the other side (thickness 150μm NexTech Materials, Ltd) or dense plain YSZ disks (of diameter 20mm and thickness 1mm). YSZ pellets were used as substrates for the deposition of cathodic films, while porous LSM disks (thickness 2mm) for the deposition of electrolytic films. Slurries of Cu-CeO_2 were made by a wet dispersion method. These were fabricated by mixing CuO and CeO_2 powders with a mixture of organic solvents. Details of this procedure are given elsewhere [7]. Painted pellets were sintered in air at 1000°C for 5 h with heating-cooling rates at 2°C /min. Films fabricated by spray pyrolysis method were sintered in air at 700°C for 4 hours (electrodic films) or 1000°C for 5 h (electrolytic films) with the same heating-cooling rates.

C. Characterization

Samples were tested by X-ray diffraction for verification of crystal phase by means of a diffractometer. Film morphology and composition were examined by SEM/EDS. Electrochemical characterization of some samples was carried out in a suitable reactor set-up in terms of steady-state I-V measurements [9].

III. Results and Discussion

Optimization of the spray pyrolysis technique focused on the operating parameters and their effect on film morphology and adhesion with the substrate surface. The following parameters were studied in detail: a) Substrate temperature b) Deposition time c) Precursor solution concentration c) post-deposition sintering temperature d) solution flow rate. Optimum values of the above parameters are shown in table 1.

The process step of sintering follows the spraying process step. The values of all the above parameters depend on the type of the film fabricated (electrode or electrolyte). In general, low sintering temperatures can be applied with this method due to the absence of starting oxide powders that are necessary in conventional fabrication methods. As a result, the fabrication cost of the cell is reduced. Electrolytic films must be dense to prevent short-circuit of the cell while the electrode has to be porous for optimum electrochemical kinetics. For every type of film it was found that the crystal phases (see Fig.2) are formed at sintering temperature 700°C in agreement with Papastergiades et al. [3]. In order to achieve dense uniform thickness for the electrolytic films, three consecutives depositions were applied followed by sintering at 1000°C for 5 hours.

SEM photographs and corresponding substrate temperature profiles of Co-CeO_2, Cu-CeO_2 and Cu-LSCM anodic electrodes are given in Fig.3.

The surface of Co-CeO_2 and Cu-LSCM showed a uniform cracks-free morphology at substrate temperatures 264°C and 306°C respectively. The obtained pattern might be due to the effect of the substrate surface morphology. Severe cracks on the surface of Cu-CeO_2 might explained by the low average substrate temperature of 171°C in combination with the fact that the spray was done intermittently to allow for temperature recovery. The Cu-LSCM surface photo belongs to a composite structure with the substrate of YSZ fabricated also by spray pyrolysis deposited onto a porous pellet of LSM disk (thickness 2 mm).

A cross section of the composite structure Cu-LSCM/YSZ/LSM is shown in Fig. 4 with the corresponding deposition substrate temperature profile. In this case, three consecutive depositions of the electrolytic electrode at average substrate temperature of 279°C were done before the final deposition of the anodic electrode. Intermediate sintering was performed after the second deposition. This procedure resulted in a dense electrolytic film (thickness 3-4 μm) and excellent adhesion between the two deposited films. For substrate temperatures <250°C or > 350°C, flow rates >35 ml/h and concentrations > 0.025 M severe surface cracks and poor adhesion were observed. For solution concentrations greater than 0.1 M, powder formation was observed on the substrate surface.

Electrochemical measurements were performed for the Co-CeO_2 electrode (Fig. 3) in a fuel cell arrangement of the type Co-CeO_2/ScSZ/LSM with the ScSZ/LSM part fabricated with conventional methods (NexTech Materials, Ltd).

In Fig. 5, I-V-P plots are shown at temperatures 750°C and 800°C. During the experiments, the anode compartment was fed with a gas mixture of 20%H_2 (balance Ar), while the cathode was exposed to atmospheric air. Figure 5 compares the performance of a cell with a SP fabricated anode with a similar cell with a conventionally (using the wet dispersion method) fabricated anode according to Musa et al. [9]. It can be seen that a significant improvement is achieved in performance with, the SP made anodic electrode by a factor of 5 in terms of maximum power and current obtained. This improvement is probably due to the better anode/electrolyte interface that is achieved by the SP technique. It cannot be attributed to the cathodic side of the cells, as the same cathode/electrolyte interface was used in both cases.

ISBN: 978-1-61804-253-8

IV. CONCLUSIONS

The spray pyrolysis technique was used for the fabrication of anodic, cathodic and electrolytic films as well as composite structures electrolyte/anode for elementary solid electrolyte fuel cell structures (SOFC).The potential to fabricate all the components of a solid oxide fuel cell in situ by SP is presented. The spray pyrolysis operating parameters were investigated in terms of substrate temperature, total ion concentration and solution flow rate.

Optimum average substrate temperatures for the production of uniform thickness surface crack-free films were of the order of 250 – 310°C while best initial total ion concentrations were 0.025M and solution flow rates 30 – 35 ml/h. Distilled water was used as solvent for all experiments. Crystal phases were found to form at post-deposition sintering temperatures of 700°C while for dense electrolyte films multiple depositions were needed with post-deposition sintering temperatures of 1000 °C for 5 h. The successful application of this technique for fabrication of a composite structure consisting of Cu-LSCM/YSZ on cathodic substrates LSM was shown. SP fabricated anodic film of Co-CeO_2 on ScSZ/LSM substrate exhibited superior electrochemical performance compared with anodic film of the same composition made by the conventional wet slurry method. Future efforts will be directed in electrochemical measurements of composite structures by spray pyrolysis with different materials and structures. Such structures may be cathode/electrolyte interface on anodic substrate (e.g. $La_{0.75}Sr_{0.25}Co_{0.2}Fe_{0.8}O_3$/CGO on Ni-CGO pellet) or double layer electrolytic films (e.g. LSCF/CGO/YSZ on Ni-YSZ pellet).

ACKNOWLEDGMENT

The authors thank the research team of Professor M. Stoukides of Aristotle University of Thessaloniki and especially Dr. V. Kyriakou for his assistance in electrochemical measurements.

REFERENCES

[1] B. C. H. Steele, "Materials science and engineering: the enabling technology for the commercialization of fuel cell systems" *J. Mater. Sci.*, vol. 36, pp. 1053-1068, 2001.

[2] G. L. Messing, S-C Zhang, G. V. Jayanthi, "Ceramic powder synthesis by spray-pyrolysis" *J. Am. Ceram. Soc.*,vol. 76, pp. 2707-2726, 1993.

[3] E. Papastergiades, S. Argyropoulos, N. Rigakis, N. E. Kiratzis, "Fabrication of ceramic electrolytic films by the method of solution aerosol thermolysis (SAT) for solid oxide fuel cells (SOFC)" *Ionics*, vol. 15, pp. 545-554, 2009.

[4] T. Setoguchi, M. Sawano, K. Eguchi and H. Arai, "Application of the stabilized zirconia thin film prepared by spray pyrolysis method to SOFC" *Solid State Ionics*, vol. 40/41, pp. 502–505, August 1990.

[5] L. Liu, G-Y. Kim, A. Chandra, "Fabrication of solid oxide fuel cell anode electrode by spray pyrolysis" *J. Power Sources*, vol. 195, pp. 7046-7053, 2010.

[6] D. Beckel, A. Dubach, A.R. Studart, L. J. Gauckler, "Spray pyrolysis of $La_{0.6}Sr_{0.4}Co_{0.2}Fe_{0.8}O_{3-\delta}$ thin film cathodes" *J. Electroceram.* Vol. 16, pp. 221-228, 2006.

[7] N. E. Kiratzis, P. Connor, J. T. S. Irvine, *J. Electroceram.*, vol. 24, pp. 270-287, 2010.

[8] D. W. Dees, T. D. Claar, T. E. Easler, D. C. Fee, F. C. Mrazek., "Conductivity of Porous Ni / ZrO_2 - Y_2O_3 Cermets", *J. Electrochem. Soc.*, vol. 134, no. 9, pp. 2141-2146, 1987.

[9] A. A. Al-Musa, V. Kyriakou, M. Al-Saleh, R. Al-Shehri, N. Kaklidis, and G. Marnellos, *ECS Trans.*, vol. 58, no.3, pp. 131-143, 2013.

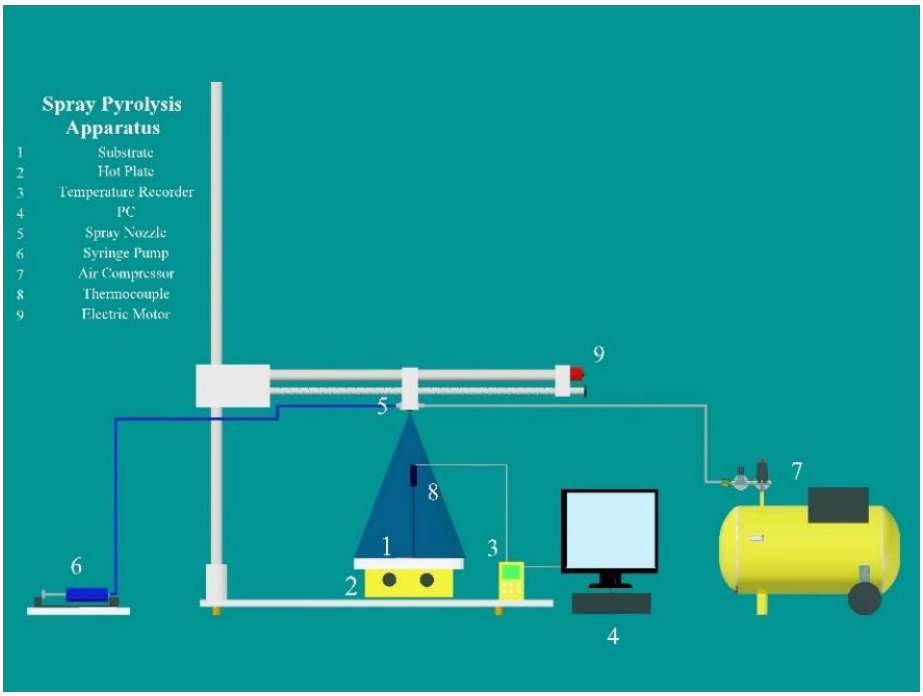


Fig. 1 Spray Pyrolysis apparatus

Table 1: Spray pyrolysis optimum operational parameters

Spray nozzle – substrate distance	20	cm
Average temperature substrate	250-310	^{0}C
Solution flow rate	30 – 35	ml/h
Air pressure	1	atm
Total ion concentration	0.025	M
Sintering temperature (electrodes)	700	^{0}C
Sintering time (electrodes)	4	Hour

ISBN: 978-1-61804-253-8

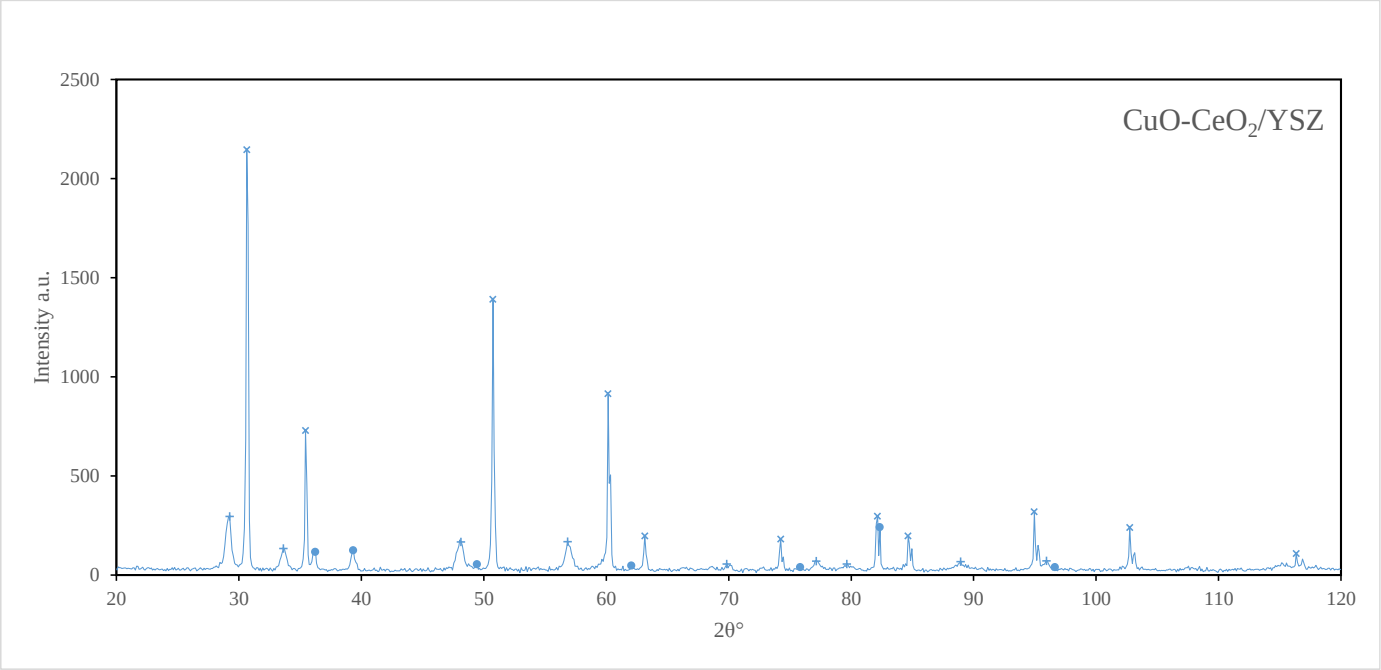


Fig.2. XRD- of anodic film (key: •CuO, $^{+}CeO_2$, ×YSZ)

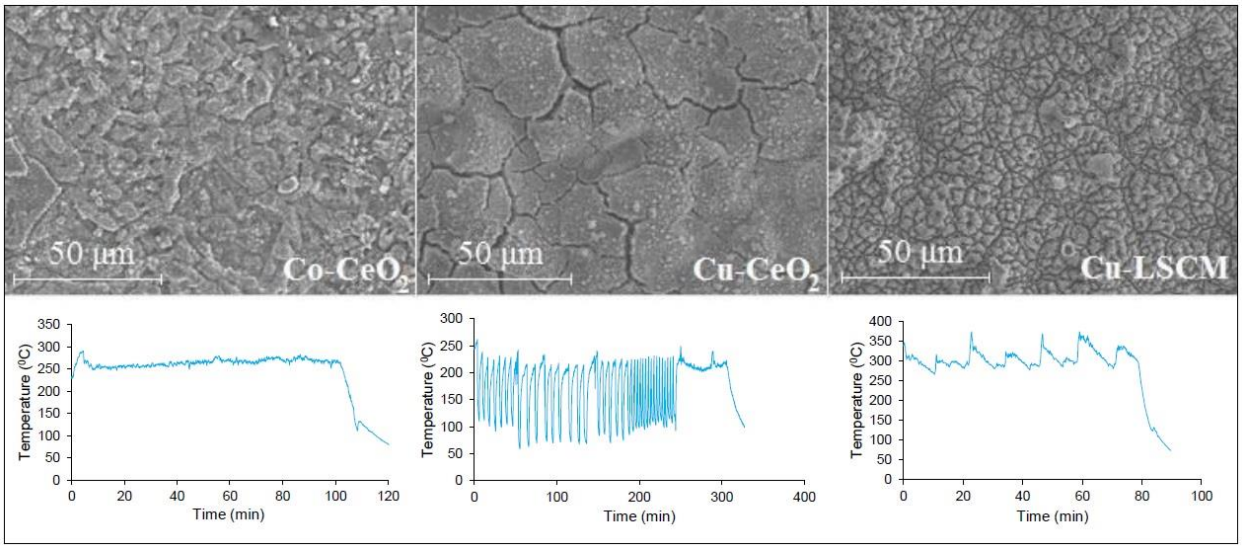


Fig 3. Anodic electrodes prepared by sp with corresponding substrate temperatures variation.

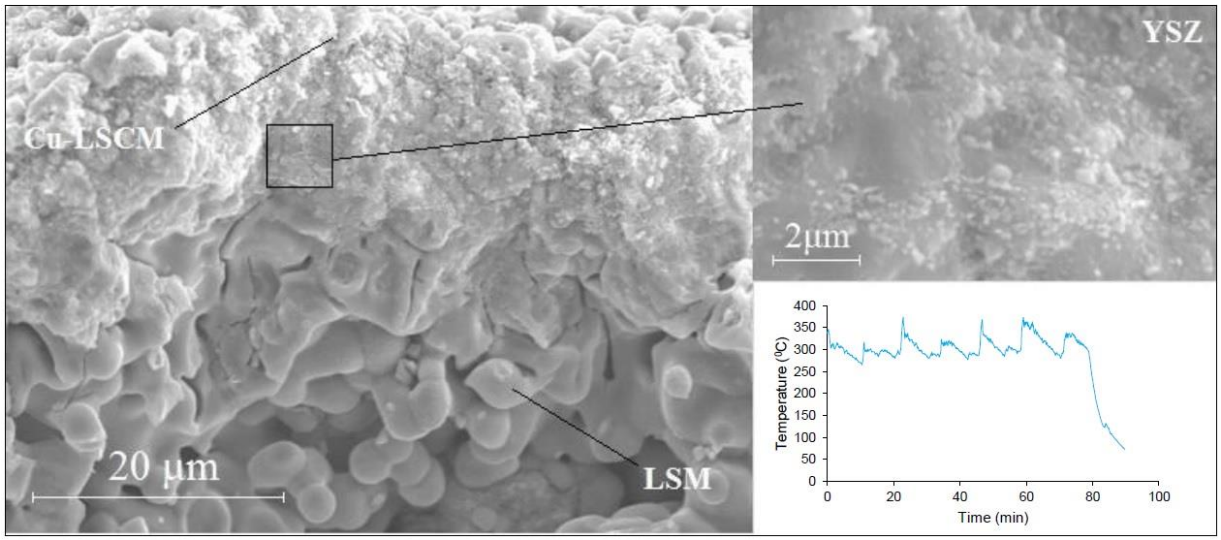


Fig 4. Cross section SEM of the composite Cu-LSCM/YSZ/LSM and substrate temperature profile of the Cu-LSCM deposition.

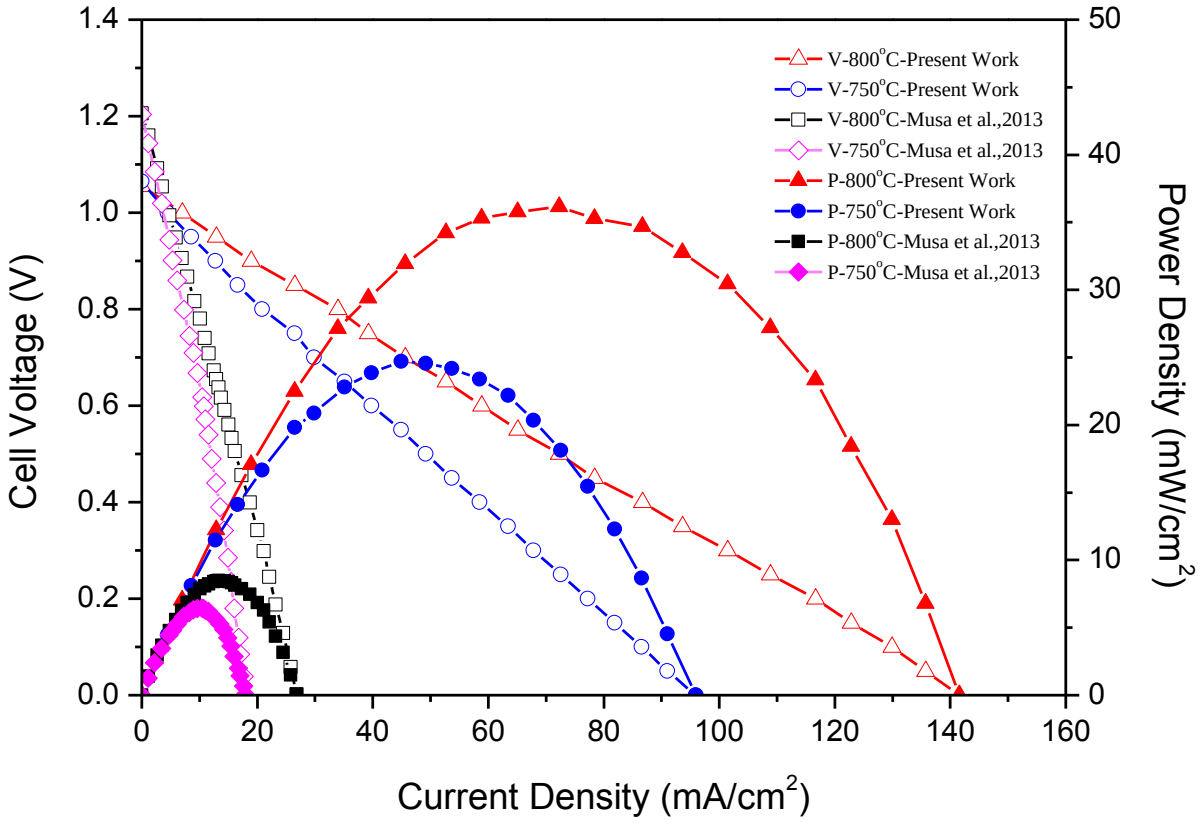


Fig 5. I-V-P Steady state results of Co-CeO2/ScSZ/LSM cells with anodes made by SP and conventional methods.

ISBN: 978-1-61804-253-8

Authors Index

ISBN: 978-1-61804-253-8