Technical Physics, Vol. 48, No. 11, 2003, pp. 1359–1363. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 1–5.

THEORETICAL AND MATHEMATICAL PHYSICS

Classical and Nonclassical Symmetries of the Nonlinear Equation with Dispersion and Dissipation

V. V. Gurskiĭ*, A. M. Samsonov*, and F. Schwarz**

** Ioffe Physicotechnical Institute, Russian Academy of Sciences, Politekhnicheskaya ul. 26, St. Petersburg, 194021 Russia*

e-mail: samsonov@math.ioffe.ru

*** Fraunhofer-Einrichtung für Algorithmen und Wissenschaftliches Rechnen (SCAI), St. Augustine, D-53754, Germany*

Received April 17, 2003

Abstract—Nonclassical symmetries of the fourth-order nonlinear partial differential equation with dispersion and dissipation are obtained and are used as a basis for deriving new exact solutions that are invariant with respect to these symmetries. The equation describes the propagation of nonlinear long-wavelength longitudinal deformations in an elastic rod placed in an external dissipative medium, the waves at the surface of a viscous liquid, etc. The solutions describing running waves are investigated based on the classical symmetries of a reduced version of the basic equation. It is shown that such solutions can be constructed within the class of elliptic functions. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The goal of the present paper is to construct new solutions to the nonlinear quasi-hyperbolic equation with dissipation that describes the propagation of deformation waves in a nonlinearly elastic rod placed in a dissipative or an active medium [1]. The new solutions are constructed based on the Lie symmetries obtained for the equation. This quasi-hyperbolic equation contains fourth-order derivatives accounting for the dispersion and a third-order derivative describing the dissipation of energy into the environment through the side surface of the waveguide. In view of the fact that this equation, which is applicable to many problems about long-wavelength waves in a waveguide, is complicated and universal, any exact solutions are very important for applications and numerical experiments. Exact solutions to differential equations that are invariant under groups of point transformations are often asymptotically stable attractors (in a certain functional norm) of solutions to the initial- and boundary-value problems for these equations.

The classical theory of Lie point symmetries for differential equations describes the groups of infinitesimal transformations in a space of dependent and independent variables that leave the manifold associated with the equation unchanged [2–4]. The contact transformations (the Lie–Baucklund symmetries) are generalizations of such symmetries and include transformations of derivatives [2, 3]. However, the class of partial differential equations (PDEs) possessing nontrivial point or contact symmetries is fairly narrow, which limits the applicability of the method of classical symmetries.

If the equation contains free parameters or arbitrary functions, then, in some cases, it is possible to impose such restrictions on them that the equation will possess a nontrivial symmetry and thereby to construct the corresponding invariant solution. Since the relevant calculations are usually rather laborious, they can be conveniently carried out by means of symbolic computations. In our work, we used the ALLTYPES computer software package, which was devised earlier by one of us [5].

Along with the method of symmetries, various methods are often used in which exact solutions to nonlinear equations are sought in a straightforward way. Some of these methods were proposed and considered in [1, 6–9]. Generalizations of the methods of point and contact classical Lie symmetries were also developed [10–12]; it was shown that these new group methods (involving nonclassical and conditional symmetries) can produce new invariant solutions. In particular, it was established that any reduction of the PDE to an ordinary differential equation (ODE) or a set of ODEs is equivalent to the existence of a nonclassical symmetry possessed by the PDE [6, 11, 13]. However, mathematically correct relationships between the method of symmetries and the methods based on direct integration of nonlinear equations are as yet undetermined.

In this paper, we will show that the nonlinear equation under analysis possesses classical and nonclassical symmetries. Nonclassical symmetries lead to a new exact solution that can describe both the effects of relaxation of a dynamic load to a static stress and the parametric oscillations of elastic stresses, whereas classical invariant solutions describe the running waves. We

1063-7842/03/4811-1359$24.00 © 2003 MAIK "Nauka/Interperiodica"

will consider several methods for constructing such solutions; in particular, we investigate the classical symmetries of the corresponding ODE for running waves. The invariant solutions generated by these symmetries are derived in terms of Jacobian elliptic functions with a fixed modulus.

CLASSICAL AND NONCLASSICAL SYMMETRIES

We consider the following nonlinear quasi-hyperbolic PDE with two dispersions and with dissipation:

$$u_{tt} - c^2 u_{xx} - (u^2 + a u_{xx} + b u_{tt})_{xx} - \mu u_{xxt} = 0. \quad (1)$$

Here, t is the time; x is the spatial coordinate along the longitudinal axis of a finite cross section rod; and a, b, c, and μ are constants dependent on the properties of the waveguide and environment. The details of the derivation of Eq. (1) can be found in [1]. In the absence of dissipation ($\mu = 0$), this equation describes the propagation of nonlinear longitudinal deformation waves with the amplitude $u(x, t)$ in a nonlinearly elastic rod and is called the equation with two dispersions. Strictly speaking, the amplitude $u(x, t)$ is the x derivative of the longitudinal displacement, i.e., the component of the displacement gradient. The first term on the right-hand side of Eq. (1) describes the elastic nonlinearity of the waveguide, and the second and third terms account for the dispersion of the waves due to the small (but finite) cross section of the rod. It is assumed that the rod is immersed in an external elastic medium and that the reaction force at the rod surface has a dissipative (active) component [1, 14]. Taking into account such a contact between the rod and the environment and performing some manipulations gives rise to the dissipative (active) term with the coefficient μ in Eq. (1) [1]. Equations analogous to Eq. (1) also arise in the theory of waves at the surface of shallow water (in which case the coefficient μ is proportional to the water viscosity) and in some other applications.

In order to find the operators of classical and nonclassical symmetries possessed by Eq. (1), we consider the vector field of infinitesimal transformations in the phase space (x, t, u):

$$X = \xi(x, t, u)\frac{\partial}{\partial x} + \eta(x, t, u)\frac{\partial}{\partial t} + \varphi(x, t, u)\frac{\partial}{\partial u}. \quad (2)$$

The functions $u = u(x, t)$, which are invariant under the infinitesimal transformations X, are, in essence, solutions to an equation arising as the "invariant surface condition":

$$\varphi(x, t, u) - \xi(x, t, u)\frac{\partial u}{\partial x} - \eta(x, t, u)\frac{\partial u}{\partial t} = 0. \quad (3)$$

This condition is a first-order PDE for $u(x, t)$. Let us consider the space J^4 with the Cartesian coordinates $(x, t, u, u^{(4)})$, where the symbol $u^{(4)}$ denotes the x and t derivatives of u up to the fourth order. In this space, Eq. (1) can be thought of as belonging to a manifold $E \subset J^4$ defined by the equalities

$$E: F(x, t, u, u^{(4)}) \equiv u_{tt} - c^2 u_{xx} - 2u_x^2 - 2uu_{xx} - a u_{xxxx} - b u_{xxtt} - \mu u_{xxt} = 0.$$

In this formula, the function F is defined as the left-hand side of Eq. (1) and all the derivatives should be understood as independent coordinates in J^4. We denote by $M \subset J^4$ the manifold of solutions to Eq. (3) that are invariant under the infinitesimal transformations X.

Vector field (2) in the above phase space generates the extended vector field $X^{(4)}$ in J^4 such that

$$X^{(4)} = X + \zeta^x \frac{\partial}{\partial u_x} + \zeta^{tt}\frac{\partial}{\partial u_{tt}} + \zeta^{xx}\frac{\partial}{\partial u_{xx}} + \zeta^{xxt}\frac{\partial}{\partial u_{xxt}} + \zeta^{ttxx}\frac{\partial}{\partial u_{ttxx}} + \zeta^{xxxx}\frac{\partial}{\partial u_{xxxx}},$$

$$\zeta^\alpha = D_\alpha\varphi - u_x D_\alpha\xi - u_t D_\alpha\eta, \quad \alpha \in \{x, t\},$$

$$\zeta^{\alpha\beta_1, \ldots, \beta_N} = D_\alpha\zeta^{\beta_1, \ldots, \beta_N} - u_{x\beta_1, \ldots, \beta_N} D_\alpha\xi - u_{t\beta_1, \ldots, \beta_N} D_\alpha\eta, \quad (4)$$

$$\alpha, \beta_1, \ldots, \beta_N \in \{x, t\}, \quad s = 1\text{–}4,$$

$$D_\alpha = \partial_\alpha + u_\alpha\frac{\partial}{\partial u} + \sum_{\beta_1} u_{\alpha\beta_1}\frac{\partial}{\partial u_{\beta_1}} + \sum_{\beta_1, \beta_2} u_{\alpha\beta_1\beta_2}\frac{\partial}{\partial u_{\beta_1\beta_2}} + \ldots,$$

$$\alpha, \beta_1, \beta_2, \ldots \in \{x, t\}.$$

The classical Lie symmetries are defined as solutions to the following set of equations (defining equations) for the components of the operator X [3, 4]:

$$(X^{(4)}F)\big|_E = 0.$$

The nonclassical symmetries include the vector fields X satisfying the following modified defining equations [12]:

$$(X^{(4)}F)\big|_{E \cap M} = 0. \quad (5)$$

The nonclassical symmetries do not preserve the entire manifold E for the equation but merely its intersection with the manifold M. This intersection contains solutions to the equation that are invariant under the infinitesimal transformations X.

Defining equations (5), which have the form of nonlinear PDEs for the functions ξ, η, and φ, cannot be solved in the general case. To obtain particular solutions, we had to make additional simplifying assumptions about the functional form of the solution to

Eq. (5). After some calculations, we succeeded in finding the following symmetries of Eq. (1):

$$X_1 = \frac{\partial}{\partial t}, \quad X_2 = \frac{\partial}{\partial x},$$
$$Y = \frac{\partial}{\partial x} + (f(t)x + h(t))\frac{\partial}{\partial u}, \tag{6}$$

where the functions $f(t)$ and $h(t)$ satisfy the equations

$$f'' = 6f^2, \quad h'' = 6fh. \tag{7}$$

The first of Eqs. (7) is the differential equation determining the Weierstrass elliptic function $\mathcal{P}$. The operators X_1 and X_2 produce classical symmetries of the equation, while the operator Y generates essentially nonclassical symmetries.

NONCLASSICAL INVARIANT SOLUTIONS

By the definition of nonclassical symmetries, the solution $u_Y(x, t)$, which is invariant under the infinitesimal transformations Y, also satisfies invariant surface condition (3) for Y. The solution to Eq. (3) has the form

$$u_Y(x, t) = \frac{f(t)}{2}x^2 + h(t)x + k(t), \tag{8}$$

where the function $k(t)$ satisfies the ODE

$$k'' - 2fk = 2h^2 + f(c^2 + 6bf) + hf'. \tag{9}$$

The solution to the equation for $f(t)$ can be written in the form

$$f(t) = \mathfrak{P}(t + \alpha; 0, g_3), \tag{10}$$

where α and g_3 are arbitrary constants.

In order to examine the solution u_Y, we simplify the function $f(t)$ by setting the third invariant of the function $\mathfrak{P}$ equal to zero, $g_3 = 0$. In this particular case, the function $f(t)$ degenerates into a rational function,

$$f(t) = \frac{1}{(t + \alpha)^2},$$

which enables the functions $h(t)$ and $k(t)$ to be calculated in explicit form:

$$h(t) = \frac{1}{(t + \alpha)^2}$$
$$\times\left(\beta + \gamma\left[\frac{t^5}{5} + \alpha t^4 + 2\alpha^2 t^3 + 2\alpha^3 t^2 + \alpha^4 t\right]\right),$$
$$k(t) = \frac{P_{10}(t, \alpha, \beta, \gamma, \delta, \lambda)}{(t + \alpha)^2} + \frac{2\mu\log(t + \alpha)}{3(t + \alpha)},$$

where β, γ, δ, and λ are arbitrary constants and P_{10} is a tenth-degree polynomial in the variable t. The coefficients of the polynomial P_{10} are determined by the constants α, β, γ, δ, and λ. The explicit expression for the polynomial is fairly involved and is not presented here.

Setting $\gamma = \lambda = 0$, we arrive at one of the solutions u_Y that are finite in time:

$$u_Y(x, t) = \frac{1}{(t + \alpha)^2}\left(\frac{x^2}{2} + \beta x\right) + k(t),$$
$$k(t) = \frac{3b + \beta^2}{2(t + \alpha)^2} + \frac{\delta}{t + \alpha} - \frac{c^2\left(t + \frac{\alpha}{3}\right)}{2(t + \alpha)} \tag{11}$$
$$+ \mu\left(\frac{2}{9(t + \alpha)} + \frac{2}{3}\frac{2\log(t + \alpha)}{t + \alpha}\right).$$

We thus see that the solution u_Y is a polynomial in the spatial variable x and, at the same time, it is a complicated function of time (see solution (10)). Solution (11), which was obtained under the assumption that $g_3 = 0$ in solution (10), describes the relaxation of the deformation u of a nonlinearly elastic waveguide to a certain constant deformation determined by the parameters in the equation in question. In addition, by appropriately choosing the displacement α in the argument of the function $\mathfrak{P}$ in solution (10), we can represent the function $f(t)$ in terms of Jacobian elliptic functions, which are finite and periodic. In this representation, the solution u_Y describes the deformation waves that are periodic in time. Note that, physically, such regimes can generally occur only under special boundary conditions.

Another interesting feature of the solution u_Y in representation (11) is that all the terms in basic nonlinear equation (1) make independent contributions to solution (11). We can see that the first fraction in the function $k(t)$ in representation (11) is determined only by the coefficient b, i.e., by the mixed fourth-order derivative in Eq. (1). The third fraction is seen to depend on the linear term with the coefficient c. Finally, the dissipation term in the equation containing μ describes the temporal dynamics determined by the last two fractions in the function k. The fact that the terms of different physical origin in the basic equation contribute additively to the solution to the nonlinear problem is rather unusual for exact solutions to nonlinear PDEs, although it is sometimes encountered in the asymptotic solutions of the problem.

CLASSICAL INVARIANT SOLUTIONS

The solutions that are left unchanged by the classical symmetries $X = X_1 - VX_2 = \partial_t - V\partial_x$ with the constant V describe running waves whose amplitudes $u = u(z)$ depend on the phase coordinate $z = x - Vt$. Under the corresponding boundary conditions, these solutions

satisfy the nonlinear second-order ODE

$$(a + bV^2)u''(z) - \mu V u'(z) + u'(z)^2 + (c^2 - V^2)u(z) = A, \tag{12}$$

where A is an arbitrary constant.

Since, in the paper by Sophus Lie [15], such dissipative ODEs were investigated probably for the first time, Eq. (12) may be called the Lie equation.

Equation (12) can be solved explicitly either by applying a rather general method described in [1] (which is based on the differential substitution and the reduction of equation to the Abelian equation) or by using a certain special ansatz. In considering the third approach, which is based on the theoretical group method of seeking possible new solutions to the dissipative equation under discussion and to analogous dissipative equations, we investigate the classical symmetries of Eq. (12) in order to find the corresponding invariant solutions.

Applying the above procedure, we can show that, under the restrictions

$$A = 0, \quad (a + bV^2)(V^2 - c^2) = \frac{6}{25}\mu^2 V^2,$$

Eq. (12) possesses the symmetries

$$X_1 = \frac{\partial}{\partial z}, \quad X_2 = \exp\left(-\frac{5}{6\mu V}(V^2 - c^2)z\right) \times \left[\frac{3\mu V}{5(V^2 - c^2)}\frac{\partial}{\partial z} + (u - V^2 + c^2)\frac{\partial}{\partial u}\right] \tag{13}$$

with the commutator $(X_1, X_2) = -[5/(6\mu V)](V^2 - c^2)X_2$. Applying this two-parameter symmetry group in the same manner as was described in [15], we can construct the solutions invariant with respect to symmetries (13):

$$u = B\left[1 - \xi^{-2}\exp\left(\frac{5B}{3\mu V}z\right)\right], \quad \int \frac{d\xi}{\sqrt{1 - C_1\xi^6}} = \exp\left(\frac{5B}{6\mu V}z\right) + C_2, \tag{14}$$

where $B = (V^2 - c^2)$ and C_1 and C_2 are arbitrary constants.

For $C_1 > 0$, the integral in solutions (14) reduces to an elliptic integral of the first kind. Inverting the latter yields the elliptic function ξ as a function of z. Substituting the elliptic function so obtained into the first of formulas (14) and performing the necessary manipulations, we arrive at two solutions:

$$u_{\pm}(z) = B\left[1 - C_1^2\exp\left(\frac{5B}{3\mu V}z\right) \times \left(1\sqrt{3} + 2\sqrt{3}\frac{1 \pm CN}{SN^2}\right)\right]. \tag{15}$$

Here, we have introduced the notation

$$SN \equiv \mathrm{sn}\left(2(3)^{1/4}C_1\exp\left(\frac{5B}{6\mu V}z\right) + C_2, k\right),$$

$$CN \equiv \mathrm{cn}\left(2(3)^{1/4}C_1\exp\left(\frac{5B}{6\mu V}z\right) + C_2, k\right),$$

where sn and cn are the first and second Jacobian elliptic functions (the elliptic sine and cosine, respectively).

The modulus of these functions is determined by the formula

$$k = \frac{\sqrt{2 - \sqrt{3}}}{2} \approx 0.26.$$

In these formulas, we have used, for simplicity, the same notation C_1 and C_2 as in solutions (14) but for the new arbitrary constants.

CONCLUSIONS

We have investigated the symmetries of nonlinear equation (1) with two dispersions and with dissipation, which may serve as a natural generalization of the equation describing long-wavelength waves in a waveguide with allowance for nonlinear, dissipation, and dispersion terms. As a method of investigation, we have utilized the method of nonclassical symmetries. This method permitted us to show that the only classical symmetries of Eq. (1) are those generated by the operators X_1 and X_2, with which Eq. (1) can be reduced to ODE (12) for running waves. On the other hand, the nonclassical approach makes it possible to obtain a new symmetry generated by the operator Y and to use it to construct invariant solution (8). We have also shown that the solutions to Eq. (1) that describe running waves can be derived as invariant solutions to the corresponding ODE, which are expressed in terms of the Jacobian elliptic functions with a fixed modulus.

The results obtained raise the hope that other possible generalizations of the theoretical group methods can extend the class of known exact solutions to physically important nonlinear equations. Thus, for some well-studied equations (e.g., the Korteweg–de Vries equation and its modified version, the Burgers equations, the nonlinear heat-conduction equation, etc.), fairly wide families of the nonclassical symmetries that they possess have already been derived (see, e.g., [12, 16] and the literature cited therein). We have found that the nonclassical method also works well for Eq. (1), which is a physically meaningful (although not so well studied) example of the nonlinear equations in question. Another possible method of investigation may consist in searching for effective ways of solving nonlinear defining equations (5) of the operators generating nonclassical symmetries.

ACKNOWLEDGMENTS

This work was supported in part by INTAS, grant no. 99-00167.

REFERENCES

1. A. M. Samsonov, *Strain Solitons in Solids and How to Construct Them* (Chapman and Hall, London, 2001).
2. N. Kh. Ibragimov, *Transformation Groups Applied to Mathematical Physics* (Nauka, Moscow, 1983; Reidel, Dordrecht, 1985).
3. P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer, New York, 1993; Mir, Moscow, 1989).
4. L. V. Ovsyannikov, *Group Analysis of Differential Equations* (Nauka, Moscow, 1978; Academic, New York, 1982).
5. F. Schwarz, Acta Appl. Math. **60**, 39 (2000).
6. F. Cariello and M. Tabor, Physica D **53**, 59 (1991).
7. P. A. Clarkson and M. D. Kruskal, J. Math. Phys. **30**, 2201 (1989).
8. V. A. Galaktionov, Proc. R. Soc. Edinburgh, Sect. A: Math. Phys. Sci. **125**, 225 (1995).
9. A. M. Samsonov, Appl. Anal. **57**, 85 (1995).
10. P. J. Olver and P. Rosenan, J. Math. Mech. **47**, 263 (1987).
11. P. J. Olver, Proc. R. Soc. London, Ser. A **444**, 509 (1994).
12. P. J. Olver and E. M. Vorob'ev, *CRC Handbook of Lie Group Analysis of Differential Equations*, Ed. by N. H. Ibragimov (CRC, Boca Raton, 1996), Vol. 3, p. 291.
13. D. J. Ariigo, P. Broadbridge, and J. M. Hill, J. Math. Phys. **34**, 4692 (1993).
14. A. D. Kerr, J. Appl. Mech. **31**, 491 (1964).
15. S. Lie, Arch. Math. **8**, 371 (1883).
16. E. M. Vorob'ev, Acta Appl. Math. **26**, 61 (1992).

Translated by O. Khadin

Technical Physics, Vol. 48, No. 11, 2003, pp. 1364–1371. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 6–14.

THEORETICAL AND MATHEMATICAL PHYSICS

New Data Processing Technique Based on the Optimal Control Theory

K. N. Kozlov* and A. M. Samsonov**

** St. Petersburg State Polytechnical University, Politekhnicheskaya ul. 29, St. Petersburg, 195251 Russia*

*** Ioffe Physicotechnical Institute, Russian Academy of Sciences, Politekhnicheskaya ul. 26, 194021 Russia*

Received May 19, 2003

Abstract—The problem of choosing the best set of parameters for a given mathematical model that adequately describes independent experimental data is formulated in terms of the optimal control theory. The sum of squares of discrepancies between experimental data and their analogues calculated within the framework of a given mathematical model of a process is minimized. A solution to the problem is found, and conditions for optimally choosing the parameters of the mathematical model are established. The search algorithm is generalized for the case where a penalty function is present, and an efficient way of including inequality constraints is suggested. The algorithm was tested by finding the thermal conductivity of single crystals (Ioffe–Ioffe classical experiment), thermal diffusivity of a thin plate, and parameters of gene expression during the fruit fly (*Drosophila melanogaster*) embryo evolution. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

Let a boundary-value problem for a set of equations with given differential operator and boundary (initial) conditions be the mathematical model of a process. In check experiments, the coefficients of the equations, as well as dissimilar terms in the equations and boundary conditions, are unknown (or partially known) and have to be found in such a way that the model describes adequately experimental data obtained independently. Usually, a set of boundary-value problem parameters (coefficients of equations, source functions, etc.) that best fits experimental data under given conditions is sought in this case. If mere interpolation fails, the problem may be solved in terms of the optimal control theory. The development of such an approach to experimental data processing is the aim of this work.

To find the phenomenological parameters of a model, one may apply the least-squares method to fit experimental data. Then, in terms of the optimal control theory, a model quality functional to be minimized is the sum of the squares of deviations of experimental data from values calculated independently within the framework of this mathematical model. The deviations are summed at times they were determined in experiments. Such a functional may have several local minima and, if necessary, a penalty function. Also, inequality constraints may be imposed on some of the problem parameters.

In the simplest statement, this problem was briefly considered in [1]. In the work cited, an associated algorithm was described and generalized for the case with a penalty function and an efficient way of taking into account inequality constraints was suggested.

1. STATEMENT OF THE PROBLEM

Suppose we know the values of some vector function $y(t_i) = (y_0(t_i), \ldots, y_{K-1}(t_i))^T$ that characterizes the state of a system at different times. The superscript T hereafter means transposition; $i = 1, \ldots, J$, where J is the total number of time instants at which independent experimental data were obtained; and K is the number of state variables for a system studied. We assume that a system of first-order differential equations in the independent variable t together with a boundary condition is given and that this system depends on the vector of parameters $q = (q_0, \ldots, q_{I-1})^T$, where I is the number of parameters:

$$\frac{\partial v}{\partial t} = f(v, q); \quad v(0) = v_0. \tag{1}$$

This system describes the behavior of a physical system in experiments. The left side of (1) is the vector of dimension K, which is composed of time derivatives of the function $v(t, q)$ (state variables), and the right side of (1) is the vector function $f(v, q)$. If necessary, a penalty function $P(q)$ can be introduced into the problem.

Let us introduce a set of indices of parameters I_l on which inequality constraints

$$q_i^{\text{low}} \le q_i \le q_i^{\text{up}}, \quad i \in I_l \subset I \tag{2}$$

are imposed.

1063-7842/03/4811-1364$24.00

Optimal control in this case consists in choosing parameters such that the quality functional (a measure of deviation of measured data from those calculated in terms of an independent model for a physical process)

$$F(v, q) = \sum_{i=1}^{J} (v(t_i, q) - y(t_i))^T (v(t_i, q) - y(t_i)) + P(q) = \varphi(v_1, \ldots, v_J) + P(q), \quad v_i = v(t_i, q) \tag{3}$$

is minimized.

Note that many problems of mathematical physics that are stated in the form of higher order differential equations can be reduced to the normal form, i.e., to a system of first-order equations with additional parametric variables (Pfaffian special system) [2]. Therefore, the selection of system (1) as the basic set of the problem stated is justified.

Finding the minimum of the quality functional necessitates the derivation of first-order stationary conditions. For system (1), which is written in general form, they are conveniently found with the Lagrange method of multipliers. In this case, however, one must consider an extended quality functional that includes inequalities (2), which impose restrictions on the control parameters. To this end, inequalities (2) must be transformed into equivalent equalities.

2. NECESSARY CONDITIONS FOR MINIMUM

To derive necessary conditions for stationarity (optimality) for quality functional (3), it is necessary to introduce additional controls u_i for which one can write equivalent equalities. Since the choice of these equalities is ambiguous, we will consider algebraic and trigonometric transformations of the inequalities into equivalent equalities.

(1) Algebraic transformation of the restricting inequalities. Let us replace inequalities (2) with the algebraic equalities [3]

$$\xi_i = (q_i - q_i^{\text{low}})(q_i^{\text{up}} - q_i) - u_i^2 = 0. \tag{4}$$

It is obvious that the condition $u_i = 0$ is satisfied if the initial parameter takes on either of the two preset extreme values, $q_i = q_i^{\text{low}}$ or $q_i = q_i^{\text{up}}$, and any $u_i \neq 0$ corresponds to an intermediate value of the control parameter: $q_i^{\text{low}} < q_i < q_i^{\text{up}}$.

Next, we introduce a vector function $\psi(t)$ of Lagrangean multipliers to include Eqs. (1) into the functional and a necessary number μ_i of Lagrangean multipliers to take into account the inequality constraints, which were transformed into (4). The extended functional of the problem can now be written in the form

$$L = \varphi(v_1, \ldots, v_J) + \sum_{i=0}^{J-1} \int_{t_i}^{t_{i+1}} \psi^T(t)\left(-\frac{\partial v}{\partial t} + f(v, q)\right)dt + \mu^T \xi + P(q), \tag{5}$$

where the vectors $\mu = \{\mu_i\}$ and $u = \{u_i\}$, as well as the vector function $\xi(q, u) = \{\xi_i(q, u)\}$, are introduced for all $i \in I_l$.

Thus, the restrictions are involved in the extended functional and it reaches an extremum simultaneously with (3). This allows us to use the standard procedure for deriving the stationarity conditions.

Having calculated the first variation of the quality functional

$$\delta F = \frac{\partial \varphi}{\partial v_1}\delta v_1 + \ldots + \frac{\partial \varphi}{\partial v_J}\delta v_J + \left(\frac{\partial P}{\partial q}\right)^T \delta q, \tag{6}$$

which is a measure of discrepancy between measured and calculated data, and the first variation of differential constraints

$$-\frac{\partial \delta v}{\partial t} + \frac{\partial f}{\partial v}\delta v + \frac{\partial f}{\partial q}\delta q = 0, \tag{7}$$

which are the equations of model (1), we can write the first variation of the Lagrangean function as

$$\delta L = \sum_{i=1}^{J} \frac{\partial \varphi}{\partial v_i}\delta v_i + \sum_{i=0}^{J-1} \int_{t_i}^{t_{i+1}} \psi^T(t)\left(-\frac{\partial \delta v}{\partial t} + \frac{\partial f}{\partial v}\delta v + \frac{\partial f}{\partial q}\delta q\right)dt + \mu^T\left(\frac{\partial \xi}{\partial q}\delta q + \frac{\partial \xi}{\partial u}\delta u\right) + \left(\frac{\partial P}{\partial q}\right)^T \delta q, \tag{8}$$

where $\partial P/\partial q$ are the components of the vector $\partial P/\partial q_i$.

Integrating (8) by parts yields

$$\delta L = \sum_{i=1}^{J} \frac{\partial \varphi}{\partial v_i}\delta v_i + \sum_{i=0}^{J-1} [\psi^T(t_i + 0)\delta v_i - \psi^T(t_{i+1} - 0)\delta v_{i+1}] + \left(\frac{\partial P}{\partial q}\right)^T \delta q + \sum_{i=0}^{J-1} \int_{t_i}^{t_{i+1}} \left[\left(\psi^T \frac{\partial f}{\partial v} + \frac{\partial \psi^T}{\partial t}\right)\delta v + \psi^T \frac{\partial f}{\partial q}\delta q\right]dt + \mu^T\left(\frac{\partial \xi}{\partial q}\delta q + \frac{\partial \xi}{\partial u}\delta u\right). \tag{9}$$

Here, $\psi(t_i + 0)$ and $\psi(t_{i+1} - 0)$ are the right- and left-hand limits of the Lagrangean multipliers $\psi(t)$ at intermediate points where experimental data are available; therefore, the values of Lagrangean multipliers (the function $\psi(t)$) vary in steps.

Below are a set of first-order minimum conditions for the quality functional [4] that use the stationarity

condition $\delta L = 0$:

$$\frac{\partial f}{\partial v}\psi + \frac{\partial \psi}{\partial t} = 0 \ \forall\ t \in [t_i, t_{i+1}); \tag{10}$$

$$\frac{\partial \varphi}{\partial v_i} - \psi(t_i - 0) + \psi(t_i + 0) = 0, \quad i = 1, \ldots, (J-1); \tag{11}$$

$$\frac{\partial \varphi}{\partial v_J} - \psi(t_J) = 0; \tag{12}$$

$$\mu_i u_i = 0, \quad i \in I_l. \tag{13}$$

This set of equations solves the problem of minimizing the discrepancy between experimental data and data calculated from a solution to the boundary-value problem.

Thus, the stationarity condition can be recast in the form of a vector equality for the Lagrange function gradient ζ:

$$\zeta(v, q) \equiv \int_{t_0}^{t_J} \psi^T \frac{\partial f}{\partial q} dt + \frac{\partial P}{\partial q} + \mu^T \frac{\partial \xi}{\partial q} = 0. \tag{14}$$

(2) Numerical solution algorithm. Let a set of I parameters q be given, a physical process be described by system (1), and it be necessary to find the coefficients of the equations such that the discrepancy between a solution to the model mathematical problem and measurements is minimal at each point of a given interval. Then, a solution algorithm for the problem stated by (1)–(3) consists of the following steps.

(1) Equation (1) is integrated.

(2) Equation (10) is integrated in reverse order, i.e., from t_J to t_0, in view of initial condition (12) and conditions (11) at those intermediate points where experimental data are available.

(3) The parameter gradient $\zeta^k = \zeta(v^k, q^k)$ is calculated by formula (14).

(4) Condition (13) for Lagrangean multipliers that correspond to algebraic constraints for additional controls u_i is satisfied as follows. If a parameter q_i in (2) meets the strict inequality, $\mu_i = 0$; otherwise, $u_i = 0$ and μ_i is selected so that a new value of the parameter q_i is allowable.

(5) New values of parameters q^{k+1} are found by the formula

$$q^{k+1} = q^k - \alpha^k \zeta^k, \tag{15}$$

where k is the number of iterations and α^k is a parameter selected so that functional (3) diminishes at each step.

Steps 1–5 are repeated until a desired calculation accuracy is achieved, for example, until the value of the functional becomes less than a preset value. A vector of parameters q^N obtained at the last step specifies a solution to the problem.

(3) Trigonometric transformation of constraints. The commonly used procedure in the optimal control theory is the replacement of control parameters q_i for which inequality constraints (2) are imposed by new controls u_i by means of trigonometric relationships, for example,

$$u_i = \alpha_i + \beta_i \sin(\gamma q_i), \tag{16}$$

where the factor γ is taken so as to improve divergence during numerical experiments and the constant α and β are determined from the upper and lower limits of the initial controls:

$$\alpha_i = (q_i^{\text{up}} + q_i^{\text{low}})/2; \quad \beta_i = (q_i^{\text{up}} - q_i^{\text{low}})/2.$$

Clearly, such a transformation is not unique; therefore, it seems reasonable to consider another finite representation:

$$u_i = \alpha_i + \beta_i \tanh(\gamma q_i). \tag{17}$$

The above transformations are applied only to those q_i that must satisfy conditions (2). Thus, instead of (1), we obtain upon rearrangements

$$\frac{\partial v}{\partial t} = \bar{f}(v, \bar{q}, u), \tag{18}$$

where $\bar{q} = \{q_i\}$ for $i \notin I_l$ and $u = \{u_i\}$ for $i \in I_l$.

Designating $\{u_i\}$ as $\{q_i\}$, where $i \in I_l$, we can write (18) in the form of (1), where $q = \{q_i\}$ and $i = 0, 1, \ldots, I-1$.

To derive necessary minimum conditions for the discrepancy functional, we write the Lagrangean

$$L = \varphi(v_1, \ldots, v_J) + \sum_{i=0}^{J-1} \int_{t_i}^{t_{i+1}} \psi^T(t)\left(-\frac{\partial v}{\partial t} + f(v, q)\right)dt + P(q), \tag{19}$$

where $\psi(t)$ is, as before, the vector function of Lagrangean multipliers.

Now, there is no need for additional multipliers that include inequality constraints into the extended functional, since they are involved in the equations.

After integrating (19) by parts, the standard derivation of necessary conditions for minimum leads us to a formula for the first variation of the Lagrangean:

$$\delta L = \sum_{i=1}^{J} \frac{\partial \varphi}{\partial v_i}\delta v_i + \sum_{i=0}^{J-1} [\psi^T(t_i + 0)\delta v_i - \psi^T(t_{i+1} - 0)\delta v_{i+1}] + \sum_{i=0}^{J-1} \int_{t_i}^{t_{i+1}} \left[\left(\psi^T \frac{\partial f}{\partial v} + \frac{\partial \psi}{\partial t}\right)\delta v + \psi^T \frac{\partial f}{\partial q}\delta q\right]dt + \left(\frac{\partial P}{\partial q}\right)^T \delta q. \tag{20}$$

By virtue of the stationarity condition $\delta L = 0$, the necessary conditions for minimum have the form

$$\frac{\partial f}{\partial v}\psi + \frac{\partial \psi}{\partial t} = 0 \quad \forall\ t \in [t_i, t_{i+1}); \tag{21}$$

$$\frac{\partial \varphi}{\partial v_i} - \psi(t_i - 0) + \psi(t_i + 0) = 0, \quad i = 1, \ldots, (J-1); \tag{22}$$

$$\frac{\partial \varphi}{\partial v_J} - \psi(t_J) = 0. \tag{23}$$

Thus, the stationarity conditions may be recast as the vector equality

$$\zeta(v, q) = 0, \tag{24}$$

where ζ is the Lagrangean gradient:

$$\zeta(v, q) = \int_{t_0}^{t_J} \psi^T \frac{\partial f}{\partial q} dt + \frac{\partial P}{\partial q}. \tag{25}$$

(4) Numerical solution algorithm for the trigonometric transformation of the constraints. Let a set of I parameters q be given, a physical process be described by system (1), and it be necessary to find the coefficients of the equations such that the discrepancy between a solution to the model mathematical problem and independent measurements is minimal at each point of a given interval. Then, a solution algorithm for the problem stated by (1)–(3) consists of the following steps.

(1) System (1) is integrated with a desired accuracy.

(2) Conjugate system (21) is integrated in reverse order, i.e., from t_J to t_0, in view of (22) and (23).

(3) The parameter gradient $\zeta^k = \zeta(v^k, q^k)$ is calculated by formula (24).

(4) New values of parameters q^{k+1} are found by the formula

$$q^{k+1} = q^k - \alpha^k \zeta^k, \tag{26}$$

where k is the number of iterations and the step α^k of the gradient method is selected so that functional (3) diminishes at each step.

Steps 1–4 are repeated until a desired calculation accuracy is achieved, for example, until the value of the functional becomes less than a preset value. A vector of parameters q^N obtained at the last step provides a solution to the problem.

Unlike Section 2.2, here there is no need for step 4, since the constraints are taken into account in the new extended functional.

3. NUMERICAL EXPERIMENTS IN HEAT CONDUCTION PROBLEMS

A simple example of the problem studied is analysis of temperature fields in a sample with the subsequent determination of the coefficients of related equations. Let us demonstrate the efficiency of our data processing method by finding the thermal conductivity (thermal diffusivity) of a sample from measured temperature values.

(1) Thermal conductivity of single crystals. To verify the method, we turn to the classical results of Ioffe and Ioffe for the thermal conductivity of single crystals [5]. As in [5], assume that experimental conditions are such that the conventional heat conduction equation

$$\frac{\partial T(\tau, x)}{\partial \tau} = \frac{k}{c}\frac{\partial^2 T}{\partial x^2} \tag{27}$$

(τ is time; x is spatial coordinate; and T, k, and c are the temperature, thermal conductivity, and specific heat of the sample) is valid.

Our goal is to find the thermal conductivity coefficient k that provides the least deviation of a solution to (27) from an experimental curve.

Ioffe and Ioffe experimented with a NaCl single crystal sandwiched in copper blocks with the same initial temperature. The lower block was immersed in a coolant or liquid air. One thermocouple measured the temperature difference $T_1 - T_2$ between the blocks; the other, the temperature T_2 of the upper block relative to room temperature T_0. The readings of both thermocouples were taken in 30-s intervals for the measurement time $\tau_m = 7$ min and tabulated (Table 1).

Table 1. Experimental data for the determination of the NaCl thermal conductivity [5]

Measurement no.	τ, s	$T_1 - T_2$	T_2
1	60	7.7	0.22
2	90	10.7	1.5
3	120	12.3	3.2
4	150	13.4	5.0
5	180	13.9	6.85
6	210	14.2	8.75
7	240	15.0	10.7
8	270	15.2	12.75
9	300	16.1	14.80
10	330	16.3	16.85
11	360	16.7	18.95
12	390	16.8	21.0
13	420	16.8	23.1

Table 2. Absolute temperature values

Measurement no.	τ, s	T_2	T_1
1	0	24	24
2	60	23.78	16.08
3	90	22.5	11.8
4	120	20.8	8.5
5	150	19	5.6
6	180	17.5	3.25
7	210	15.25	1.05
8	240	13.3	–1.7
9	270	11.25	–3.95
10	300	9.2	–6.9
11	330	7.15	–9.15
12	360	5.05	–11.65
13	390	3	–13.80
14	420	0.9	–15.9

When using Table 1 [5], it is necessary to make two assumptions. First, since in the experiment provision was made for preventing heat losses, we may consider the problem as one-dimensional and direct the x axis normally to the block–sample interfaces from the warm to cold block, i.e., from top to bottom. Second, since the temperature gradient in the experiment varies slowly, we may linearly interpolate in time the temperature values at the points $x = 0$ and L, where the measurements were taken. The point $x = 0$ lies at the warm block–sample interface; the point $x = L$ ($L = 0.64$ cm), at the sample–cold block interface.

The absolute temperature values are listed in Table 2 (the initial temperature was $T_0 = 24$°C).

We discretize Eq. (27) over space, introduce a uniform mesh $\{x_i = ih, h = L(K-1)\}$ (where $i = 0, \ldots, K-1$), take into consideration that the temperature distribution depends parametrically on the thermal conductivity, and obtain the set of differential–difference equations

$$\frac{\partial T(\tau, x_i, k)}{\partial \tau} = \frac{k}{c}(T(\tau, x_{i-1}, k) - 2T(\tau, x_i, k) + T(\tau, x_{i+1}, k)); \quad i = 1, \ldots, K-2 \tag{28}$$

with the initial condition

$$T(0, x_i, k) = T_0; \quad i = 0, \ldots, K-1 \tag{29}$$

and boundary conditions

$$T(\tau, x_0, k) = T_2(\tau); \quad \forall \tau \in [0, \tau_m]; \tag{30}$$

$$T(\tau, x_{K-1}, k) = T_1(\tau); \quad \forall \tau \in [0, \tau_m]. \tag{31}$$

The only unknown parameter in set (28) is the thermal conductivity k.

The final sum of squares of discrepancies between temperatures calculated (by model (27)) and measured at all time points,

$$F(k) = \sum_{i=1}^{K-2} (T(\tau_m, x_i, k) - \bar{T}(\tau_m, x_i))^2, \tag{32}$$

gives the quality functional for the problem of experimental data processing.

Unfortunately, the function $\bar{T}(\tau_m, x)$ [5] is known only at the points $x = 0$ and L. Since the boundary conditions are approximated linearly, we may assume that the second derivative on the left of (27) is constant and, hence, the function is quadratic in x. For the function $\bar{T}(\tau_m, x)$, we took a parabola $ax^2 + bx$, since only two coefficients can be found by two points.

Calculation was performed for 40 randomly selected initial approximations that were uniformly distributed over the closed interval [0.005, 0.025]. The outcome exceeded our expectations. The average calculated thermal conductivity was $k = 0.012$, rather than $k = 0.013$, which was previously obtained by Ioffe. At the end of the calculation, the integral deviation of the calculated data from the quality functional (measured data) was found to be 9×10^{-3}.

Figure 1 shows a typical curve of convergence in the problem of thermal conductivity of single crystals.

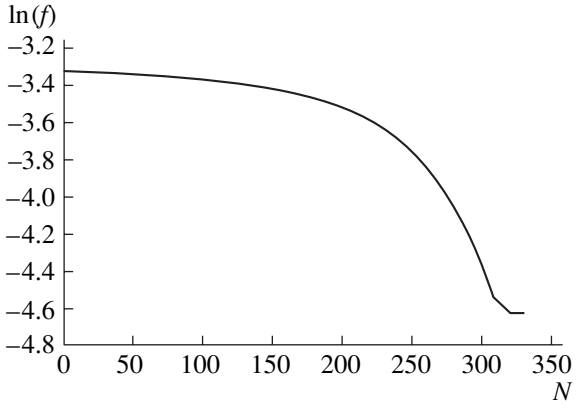


Fig. 1. Convergence for the quality functional f vs. number N of iterations upon calculating the thermal conductivity.

(2) Thermal diffusivity of a plate. By way of second example, let us consider the problem of thermal diffusivity [6] of a plate of thickness $h = 10$ cm. On one side of the plate, the temperature was T_0; on the other side and at the middle of the plate, the temperature oscillation amplitudes were $\Delta T_1 = 10$°C and $\Delta T_2 = 8$°C, respectively. The oscillation period on the other side was $\tau_0 = 1$ h.

In [6], the thermal diffusivity value was found by plotting the Fourier test and turned out to be $a = 0.009$ m^2/h.

Let us formalize the statement of the problem to apply our method. The temperature variation is described by the equation

$$\frac{\partial T(\tau, x)}{\partial \tau} = a\frac{\partial^2 T}{\partial x^2}, \quad (33)$$

where τ is time; x is the spatial coordinate; and T and a are the sample temperature and thermal diffusivity, respectively.

We discretize Eq. (33) over the coordinate x, introduce a uniform mesh $\{x_i = ir, r = h/(K-1)\}$ (where $i = 0, \ldots, K-1$), take into consideration that the temperature distribution depends parametrically on the thermal diffusivity, and obtain the set of differential–difference equations

$$\frac{\partial T(\tau, x_i, a)}{\partial \tau} = a(T(\tau, x_{i-1}, a) - 2T(\tau, x_i, a) + T(\tau, x_{i+1}, a)), \quad i = 1, \ldots, K-2 \quad (34)$$

with the initial condition

$$T(0, x_i, a) = T_0; \quad i = 0, \ldots, K-1 \quad (35)$$

and boundary conditions

$$T(\tau, x_0, a) = T_0 + \Delta T_1 \sin\left(2\pi\frac{\tau}{\tau_0}\right); \quad \forall \tau \in [0, \tau_m]; \quad (36)$$

$$T(\tau, x_{K-1}, a) = T_0; \quad \forall \tau \in [0, \tau_m]. \quad (37)$$

The only unknown parameter in set (34) is the thermal diffusivity a.

In order to write the quality functional, we must know the time dependence of the temperature in the middle of the sample, $\bar{T}(\tau, h/2)$. The period of temperature oscillation at this place is unknown, and we cannot take it to be equal to τ_0, since the thermal diffusivity would indefinitely increase in this case. Let the oscillation period at the middle be $\tau_{h/2} = 2\tau_0$ and let the measurement time be $\tau_m = \tau_0/2$. Bearing in mind that the plate is thin and the temperature conditions are quasi-steady-state [6], we linearly interpolate temperature values between the points $x = 0$, $h/2$, and h and designate the temperature distribution thus obtained as $\bar{T}(\tau_m, x)$. Then, the quality functional has the form

$$F(a) = \sum_{i=1}^{K-2} (T(\tau_m, x_i, a) - \bar{T}(\tau_m, x_i))^2. \quad (38)$$

The results are in good agreement with those obtained in [6]. Calculation was made for 20 randomly selected initial approximations that were uniformly distributed over the segment [0.0006, 0.0180]. The mean value of the thermal diffusivity was found to be $a = 0.0094$ versus $a = 0090$ in [6]. The final value of the quality functional was 23.63.

4. NUMERICAL EXPERIMENTS ON MOLECULAR BIOLOGY DATA PROCESSING

A comparatively new area of application of the algorithm suggested is the processing of huge data arrays, for example, simultaneous search for many hundreds of phenomenological parameters in mathematical problems of biology and genetics, in particular, in the gene chain model. Protein concentration greatly varies upon the evolution of a biological object and is measured with high accuracy, but finding a correlation of these data with any of the advanced theoretical models is a challenge. In one of them, the dynamics of a system is described by nonlinear diffusion reaction equations (NDREs), which contain 50 or more parameters on most of which inequality constraints are imposed.

The molecular biology problem of segment determination in a standard biological system, the fruit fly (*Drosophila melanogaster*) embryo, was stated in detail in [1]. The mathematical statement of the problem involves the set of NDREs

$$\frac{\partial v_i^a}{\partial t} = R^a g(w_i^a) + D^a[(v_{i+1}^a - v_i^a) + (v_{i-1}^a - v_i^a)] - \lambda^a v_i^a, \quad (39)$$

where the argument of g is

$$w_i^a = \sum_{j=0}^{N-1} T^{aj} v_i^j + m^a v_i^{\mathrm{bcd}} + h^a.$$

The basic element of the gene chain model is the matrix T, which characterizes a gene chain. Its elements T^{aj} describe the interplay between the concentration of one protein (the product of one gene) and the concentration of another protein (the product of another gene) by using a particular number specific for any pair of a and j.

The control actions of the protein that is the Bicoid product of maternal gene *bcd* are taken into account as an external perturbation (it is specified by the parameter m^a), and the parameter h^a reflects the threshold value of regulatory action.

Let the position of a nucleus on the longitudinal axis of the embryo be defined by index i and let all nuclei contain a copy of a regular chain of N genes, which is characterized by an $N \times N$ matrix T.

The first term on the right of (39) stands for gene regulation and protein synthesis, with the parameter R^a defining the level of synthesis. The second term describes gene product (protein) exchange between neighboring nuclei, with the parameter D^a standing for the exchange rate. The parameter λ^a describes the half-life period of the protein; the third term, gene product cleavage.

Differential constraints for the function f in (1) are given by a set of nonlinear equations like (39). Numerical simulation is aimed at determining $N(N + 5)$

Table 3. Results of calculations (see text)

Value of functional	2	3	4	5	6
<1	2(7229)	–	–	–	100(512621)
<10	13(48840)	3(77896)	–	3(449563)	–
<50	41(46000)	9(31755)	21(56027)	27(96432)	–
<100	50(52341)	9(31755)	39(54134)	33(123410)	–
>100	50(69730)	91(69873)	61(98863)	67(301301)	–

parameters of the system, where N is the number of genes.

Observables are patterns of gene activity, which supposedly are solutions to equations like (39). A desired set of parameters is that providing the closest fit of these solutions to experimental data. As before, it is necessary to minimize the sum of squares of discrepancies between the concentrations of all proteins that were calculated by the gene chain model and found in independent experiments. The sum is taken over all nuclei and all time instants for which experimental data are available.

(1) Test for efficiency of the method. Let us apply our method to data processing and finding the phenomenological parameters of a model described by a set of several hundreds of NDREs (like model (39)).

Consider a system that describes a regulatory chain of only two genes. In this case, we have 16 equations with $2(2 + 5) = 14$ parameters.

We take a certain set of parameters and solve the direct problem, i.e., integrate system (1). Let known values of the function $y(t_i) = (y_0(t_i), \ldots, y_{K-1}(t_i))^T$ be quantities $v(t_i, q) = (v_0(t_i, q), \ldots, y_{K-1}(t_i, q))^T$ that were calculated for a given number J of points over a given time interval ($i = 1, \ldots, J$) (so-called artificial data). Now, with these data, we will try to recover the initial set of parameters. In this case, the quality functional at the point of global maximum is known and equals zero.

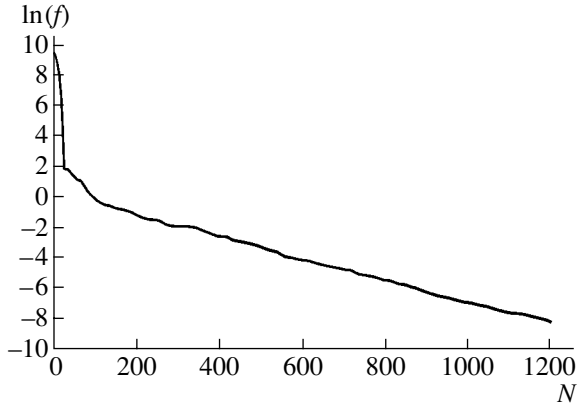


Fig. 2. The same as in Fig. 1 upon determining phenomenological parameters in the problem of molecular biology.

This allows us to estimate the quality of a solution from the value of the functional.

Let inequality constraints be imposed on six parameters linearly entering into set (39). We introduce an exponentially increasing penalty function that depends on a complex consisting of the remaining eight parameters and solve the problem of minimization numerically in the same way as before.

Calculation was performed for 100 randomly selected initial approximations that were uniformly distributed in the parameter space. For each of the initial points, we obtained several results by the method of simulated annealing [7] and the gradient method. In the latter, inequality constraints were transformed algebraically and trigonometrically (sin-transformation and tanh-transformation with various γ).

Calculation was terminated if the functional changed by less than a specified quantity at each step from a given number of steps. The set of parameters q obtained at the last step was taken to be optimal.

The numerical results listed in Table 3 are totally consistent with the theory. Columns 2–6 show the percentage of trials (initial points) when the functional falls into the interval indicated in column 1. The parenthesized figures mean the average number of integrations of the set of equations or, in other words, machine time costs. Column 6 shows the values obtained by the random search (simulated annealing) method; column 5, by the gradient method with algebraic transformation of constraints; and columns 3 and 4, by the gradient method with tanh- and sin-trigonometric transformations, respectively.

The optimal point and the number of steps turned out to be dependent on the initial approximation. For each of the initial points, calculations with tanh-transformations with various arbitrary coefficients γ were performed and the best result for each of the points was taken. These calculations are summarized in column 2.

The basic conclusions following from the calculation results are as follows.

The modified gradient method may be used to advantage in searching for the quality functional minimum when large data arrays are processed.

With the initial point selected appropriately, the number of steps needed for this method to become con-

vergent is two orders of magnitude smaller than in the random search method.

The most efficient transformation of inequality constraints is impossible to choose in advance; however, the proper selection of the parameter of tanh-transformation provided the highest rate of convergence (see Fig. 2).

DISCUSSION

Our numerical experiments show that the suggested method of processing large data arrays for finding the most appropriate parameters of a given mathematical model has a number of advantages. Namely, it offers a high rate of convergence and is applicable even if experimental data are not independent, for example, if the elements of the vector $y(t_i)$ ($i = 1, \ldots, J$) represent a Markovian sequence. In this case, the function $\varphi(v(t_1, q), \ldots, v(t_J, q))$ in (3) takes the form

$$\sum_{i=1}^{J} \ln\left[\frac{1}{\sigma_1(t_i)\sqrt{2\pi}}\exp\left\{-\frac{[y_1(t_i) - v_1(t_i, q)]^2}{2\sigma_1^2(t_i)}\right\}\right]$$

$$+ \ln\left[\frac{1}{\sigma_2(t_i)\sqrt{2\pi(1-\rho^2(t_i))}}\exp\left\{-\frac{1}{2\sigma_2^2(t_i)(1-\rho^2(t_i))}\right.\right.$$

$$\left.\left.\times\left[y_2(t_i) - v_2(t_i, q) - \frac{\rho(t_i)\sigma_2(t_i)}{\sigma_1(t_i)}(y_1(t_i) - v_1(t_i, q))\right]^2\right\}\right] \quad (40)$$

$$+ \ln\left[\frac{1}{\sigma_K(t_i)\sqrt{2\pi(1-\rho^2(t_i))}}\exp\left\{-\frac{1}{2\sigma_K^2(t_i)(1-\rho^2(t_i))}\right.\right.$$

$$\left.\left.\times\left[y_K(t_i) - v_K(t_i, q) - \frac{\rho(t_i)\sigma_K(t_i)}{\sigma_{K-1}(t_i)}\right.\right.\right.$$

$$\left.\left.\left.\times(y_{K-1}(t_i) - v_{K-1}(t_i, q))\right]^2\right\}\right].$$

In [7, 8], a version of the random search method (the method of simulated annealing or the Metropolis method) was applied to find phenomenological parameters that provide a global minimum of the quality functional in a similar problem of mathematical biology. It was shown that the given functional has many local minima but simulated annealing "bypasses" them, finding a global minimum.

However, the use of simulated annealing necessitates multiple integration of a set of strong nonlinear differential–difference diffusion reaction equations. Note for comparison that, in the case of the random search method, finding each subsequent approximation to the entire vector of parameters requires that the quality functional be calculated as many times as the number of the parameters involved, while in the gradient method, only once. This fact may become of special significance when a large number of proteins (the products of gene activity) is considered, since the number of integrations necessary for random search grows as the number of proteins squared.

REFERENCES

1. K. N. Kozlov, L. V. Petukhov, M. G. Samsonova, and A. M. Samsonov, Tr. St. Peterb. Gos. Tekh. Univ., Prikl. Mat. **485**, 73 (2002).
2. K. A. Lurie, *Optimal Control in the Problems of Mathematical Physics* (Nauka, Moscow, 1975).
3. V. A. Troitskiĭ and L. V. Petukhov, *Shape Optimization of Elastic Solids* (Nauka, Moscow, 1982).
4. A. I. Braison and Yu-Shi Kho, *The Applied Theory of Optimal Control* (Mir, Moscow, 1972).
5. A. F. Ioffe and A. V. Ioffe, Zh. Tekh. Fiz. **22**, 2005 (1952).
6. A. I. Pekhoviya and V. M. Zhidkikh, *Thermal Analysis of Solids* (Énergiya, Leningrad, 1968), p. 196.
7. J. Reinitz and D. Sharp, Mech. Dev. **49**, 133 (1995).
8. D. Sharp and J. Reinitz, Biosystems **47**, 79 (1998).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1372–1379. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 15–21.

THEORETICAL AND MATHEMATICAL PHYSICS

Dispersion Properties of a Periodic Semiconductor Structure in a Magnetic Field Directed along the Periodicity Axis

A. A. Bulgakov and V. K. Kononenko

Usikov Institute of Radiophysics and Electronics, National Academy of Sciences of Ukraine, 61085 Kharkov, Ukraine

e-mail: bulgakov@ire.kharkov.ua

Received September 10, 2002; in final form, April 4, 2003

Abstract—The band spectrum of natural waves in a periodic structure formed by alternating dielectric and semiconductor layers is investigated for the propagation of waves at some angle with respect to a magnetic field applied along the periodicity axis. A method is presented for deriving a dispersion equation, and its properties are analyzed numerically. It is shown that, in the absence of dissipation, there are two independent spectra of natural waves in the structure being considered, and the regions of existence of various types of such waves are classified. It is established that the transmission bands of the two spectra may supplement one another or overlap. It is found that, for the chosen magnetic-field direction, there exist numerous bands of transmission of cyclotron waves. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

It is well known [1, 2] that periodic layered media may be considered as a new type of artificial material whose properties can be efficiently controlled. The reaction of such media to electromagnetic radiation depends on the electrophysical parameters and thicknesses of layers forming the periodic structure in question, as well as on the applied magnetic field if use is made of semiconductors. In the last case, a solution to the problem of finding the band spectrum of natural waves in a periodic structure is determined by its configuration and by the direction of wave propagation. The configuration of the structure is specified by the direction of its periodicity and by the orientation of the applied magnetic field. For some cases that can be realized in such structures, the results of investigations can be found in the literature.

A structure where the direction of wave propagation, the direction of the applied magnetic field, and the periodicity direction coincide was considered in [3]. We note that, in a homogeneous gyrotropic medium, circularly polarized waves are normal waves, longitudinal field components being equal to zero [4]. An experimental investigation of some special features of wave propagation in such a configuration was performed in [5].

The case where the direction of the applied magnetic field is orthogonal to the periodicity direction and the plane of wave propagation was explored in [6–8]. Two waves polarized in orthogonal planes, each of these waves having three field components, are normal waves of a gyrotropic medium in this case.

In the present study, we consider a situation that is similar to that in [3] (that is, the direction of periodicity of the structure coincides with the magnetic-field direction), but we examine the case of wave propagation at some angle with respect to the magnetic-field direction (Fig. 1).

A structure whose configuration is identical to that in the present study was investigated in [9], but, there, a plasma-like layer was a two-dimensional electron gas; therefore, its thickness was made to tend to zero. In our case, a semiconductor layer has a finite thickness that is commensurate with the wavelength of the electromagnetic wave being considered. As will be shown below, this leads to a number of special physical features in the formation of the spectrum of the periodic structure.

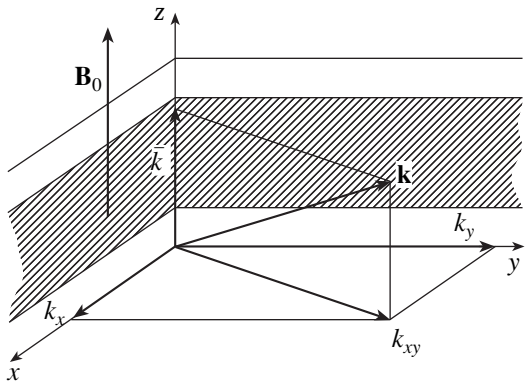


Fig. 1. Configuration of a periodic structure. Here, B_0 is the magnetic-field induction; the z axis is aligned with the periodicity direction; $\bar{k}$ is the Bloch wave number; and $\mathbf{k}(k_x, k_y, \bar{k})$ is the wave vector.

1063-7842/03/4811-1372$24.00 © 2003 MAIK "Nauka/Interperiodica"

A feature that is particular to the problem considered here and which distinguishes this problem from those that were mentioned above is that, in the present case, the Maxwell equations for a semiconductor do not separate in field components into two independent polarizations, elliptically polarized waves, each involving six field components, being normal waves of the medium in this case [4, 8].

We will study the band spectrum of natural waves in a periodic semiconductor–dielectric structure of configuration described above. The semiconductor layer of the structure is considered in the cold-plasma approximation. In a hydrodynamic description, this layer is characterized by the dielectric-permittivity tensor of the form [10]

$$\hat{\varepsilon} = \begin{vmatrix} \varepsilon_1 & i\varepsilon_2 & 0 \\ -i\varepsilon_2 & \varepsilon_1 & 0 \\ 0 & 0 & \varepsilon_2 \end{vmatrix}, \quad (1)$$

where

$$\varepsilon_1 = \varepsilon_L\left[1 + \frac{\omega_P^2}{\omega_H^2 - \omega^2}\right]; \quad \varepsilon_2 = \frac{\varepsilon_L\omega_P^2\omega_H}{\omega(\omega_H^2 - \omega^2)};$$

$$\varepsilon_3 = \varepsilon_L\left(1 - \frac{\omega_P^2}{\omega^2}\right);$$

ε_L is the dielectric permittivity of the semiconductor lattice; ω_P and ω_H are, respectively, the plasma and the cyclotron frequency; and ω is the frequency of the electromagnetic field.

The boundary conditions for the problem to be addressed assume the continuity of the tangential components of the electric and the magnetic field at the layer boundaries.

METHOD OF SOLUTION

In order to derive a dispersion equation that describes the properties of the periodic structure being considered, we will employ the transformation-matrix method [1], which makes it possible to reduce the order of the set of equations by a factor of 2.

For composing the transformation matrix in one layer, one needs independent expressions for each field component that is transverse with respect to the magnetic-field direction (in our case, these are E_x, E_y, H_x, and H_y). We note that it is these components that appear in the boundary conditions. For a gyrotropic medium, the procedure for their determination is given in [11]. It consists in introducing a scalar function $\Psi = Z(z)\psi(x, y)$ in terms of which all field components are expressed by means of simple differentiation operations. Assuming that the field components depend on the coordinate z in the direction of a constant magnetic field according to a harmonic law, we represent this function in the form

$$\Psi = (A_1\cos k_{z1}z + A_2\sin k_{z1}z + A_3\cos k_{z2}z + A_4\sin k_{z2}z)\exp[i(k_x x + k_y y)], \quad (2)$$

where k_{z1} and k_{z2} are the transverse wave numbers in a semiconductor layer. An expression that determines their values is found from the dispersion equation for a cold magnetoactive plasma [10]; for the configuration considered here, this expression takes the form

$$k_{z1,2}^2 = \frac{1}{2}\left[2\varepsilon_1 k_0^2 - \left(1 + \frac{\varepsilon_1}{\varepsilon_3}\right)k_{xy}^2\right] \pm \sqrt{\frac{1}{4}\left[2\varepsilon_1 k_0^2 - \left(1 + \frac{\varepsilon_1}{\varepsilon_3}\right)k_{xy}^2\right]^2 - \varepsilon_1\left[\varepsilon_V k_0^4 + \frac{k_{xy}^4}{\varepsilon_3} - \left(1 + \frac{\varepsilon_V}{\varepsilon_3}\right)k_{xy}^2 k_0^2\right]}, \quad (3)$$

where

$$k_0 = \frac{\omega}{c}, \quad \varepsilon_V = \varepsilon_1 - \frac{\varepsilon_2^2}{\varepsilon_1}, \quad k_{xy}^2 = k_x^2 + k_y^2.$$

Substituting the function Ψ into the corresponding expressions for the fields [11], we obtain

$$\begin{aligned} E_x &= C_1A_1\cos k_{z1}z + C_1A_2\sin k_{z1}z + C_2A_3\cos k_{z2}z + C_2A_4\sin k_{z2}z, \\ E_y &= C_3A_1\cos k_{z1}z + C_3A_2\sin k_{z1}z + C_4A_3\cos k_{z2}z + C_4A_4\sin k_{z2}z, \\ H_x &= -D_1A_1\sin k_{z1}z + D_1A_2\cos k_{z1}z - D_2A_3\sin k_{z2}z + D_2A_4\cos k_{z2}z, \\ H_y &= -D_3A_1\sin k_{z1}z + D_3A_2\cos k_{z1}z - D_4A_3\sin k_{z2}z + D_4A_4\cos k_{z2}z. \end{aligned} \quad (4)$$

The coefficients C_i and D_i are presented in Appendix 1, while the coefficients A_i can be expressed in terms of the values of the fields at the point $z = 0$. For the fields in a semiconductor layer, we eventually obtain the matrix relation

$$\begin{pmatrix} E_x(z) \\ E_y(z) \\ H_x(z) \\ H_y(z) \end{pmatrix} = \hat{S}\begin{pmatrix} E_x(0) \\ E_y(0) \\ H_x(0) \\ H_y(0) \end{pmatrix}, \quad (5)$$

where $\hat{S}$ is a 4×4 square matrix that relates the fields at an arbitrary point z of the layer to the fields at $z = 0$. The elements of the matrix $\hat{S}$ are given in Appendix 1.

For a semiconductor, the transformation matrix $\hat{S}^{-1}$ relating the fields at the beginning of a layer to those at its end can be obtained by inverting the matrix $\hat{S}$ at $z = d_1$, where d_1 is the thickness of a semiconductor layer in the structure being considered.

Similar manipulations must also be performed for a dielectric layer upon preliminarily obtaining, from Maxwell equations, independent expressions for the field components. The wave number k_z characterizing the transverse distribution of the field in a medium of dielectric permittivity ε can be obtained from (3) upon making the substitutions $\varepsilon_1 = \varepsilon_3 = \varepsilon_V = \varepsilon$ and $\varepsilon_2 = 0$. This yields

$$k_z^2 = \varepsilon k_0^2 - k_{xy}^2. \tag{6}$$

For the field components at an arbitrary point z of a dielectric layer that are expressed in terms of the field at the beginning of this layer, we have

$$\begin{aligned} E_x(z) &= F_1 E_x(0) + \frac{F_2}{\varepsilon} H_x(0) - \frac{F_3}{\varepsilon} H_y(0), \\ E_y(z) &= F_1 E_y(0) + \frac{F_4}{\varepsilon} H_x(0) - \frac{F_2}{\varepsilon} H_y(0), \\ H_x(z) &= -F_2 E_x(0) + F_3 E_y(0) + F_1 H_x(0), \\ H_y(z) &= -F_4 E_x(0) + F_2 E_y(0) + F_1 H_y(0), \end{aligned} \tag{7}$$

where the coefficients F_i are given in Appendix 2.

We can see that the matrix expression for the fields in a dielectric material has the same form as relation (5), but the matrix $\hat{S}$ is replaced there by the matrix $\hat{D}$, whose explicit form is given in Appendix 2. For a dielectric layer, the transformation matrix $\hat{D}^{-1}$ is obtained by inverting the matrix $\hat{D}$ at $z = d_2$.

Following [1] and using the condition that the fields are continuous at the boundaries of the layers, we find that the transformation matrix for one period of the structure is equal to the product of the transformation matrices for the respective superconductor and dielectric layers, $\hat{P} = \hat{S}^{-1}\hat{D}^{-1}$. This matrix relates the values of the fields at the beginning of one period of the structure to their values at its end. We note that the matrices $\hat{S}^{-1}$, $\hat{D}^{-1}$, and $\hat{P}$ are unimodular [12].

The condition of periodicity of the structure now makes it possible to express the fields at $z = 0$ in terms of the elements of the transformation matrix $\hat{P}$ for one period and the phase factor $\exp(-i\bar{k}d)$, by which, according to the Floquet (Bloch) theorem, the fields at the boundary of the period at $z = 0$ and $z = d_1 + d_2 = d$, d being the period of the structure, may differ. Specifically, we have

$$\begin{pmatrix} E_x(0) \\ E_y(0) \\ H_x(0) \\ H_y(0) \end{pmatrix} = \hat{P} \begin{pmatrix} E_x(d) \\ E_y(d) \\ H_x(d) \\ H_y(d) \end{pmatrix} = \hat{P} e^{-i\bar{k}d} \begin{pmatrix} E_x(0) \\ E_y(0) \\ H_x(0) \\ H_y(0) \end{pmatrix}, \tag{8}$$

where $\bar{k}$ is the so-called Bloch wave number—that is, the new (instead of k_{z1}, k_{z2}, k_z) period-averaged wave number of the periodic structure.

Relations (8) can be represented in the form of a set of linear homogeneous equations, and this set of equations has nontrivial solutions if and only if its determinant vanishes; that is,

$$\det(\hat{P} - e^{i\bar{k}d}\hat{E}) = 0, \tag{9}$$

where $\hat{E}$ is the identity matrix.

By explicitly evaluating the determinant in (9), we obtain the dispersion equation that relates ω, $\bar{k}$, k_x, and k_y.

INVESTIGATION OF THE DISPERSION EQUATION

We note that the dispersion equation derived from (9) is an equation of fourth degree with respect to the quantity $\xi = \exp(i\bar{k}d)$; that is,

$$\xi^4 + B_3\xi^3 + B_2\xi^2 + B_1\xi + B_0 = 0, \tag{10}$$

where the coefficients B_0, B_1, B_2, and B_3 are expressed in terms of various combinations of the elements of the matrix $\hat{P}$, the free term of the dispersion equation being equal to the determinant of the matrix $\hat{P}$; that is, $\det\hat{P} = 1$. Thus, Eq. (10) reduces to the form

$$\xi^4 + B_3\xi^3 + B_2\xi^2 + B_1\xi + 1 = 0. \tag{11}$$

The explicit expressions for the coefficients B_0, B_1, B_2, and B_3 are given in Appendix 3.

In a dispersion equation of this type, the left-hand side can be recast into the form of the product of two polynomials of second degree:

$$(\xi^2 + a_1\xi + 1)(\xi^2 + a_2\xi + 1) = 0. \tag{12}$$

This becomes possible under the condition that $B_1 = B_3$. It can be demonstrated numerically that this condition is satisfied for the configuration considered here.

It should be noted that the transition from Eq. (10) to the form (12) is a result of a numerical analysis; therefore, this result is not general. However, physics considerations suggest that, in a periodic structure, the

physical properties of the layers that constitute it must also be reflected. It follows that, if the transverse wave numbers can be found explicitly (for example, in a gyrotropic medium), it is natural to assume that, owing to the symmetry of the problem, the Bloch wave numbers will be determined by quadratic (or biquadratic) equations.

The coefficients in Eqs. (11) and (12) are related by the equations $B_1 = a_1 + a_2$ and $B_2 = 2 + a_1a_2$, whence it follows that

$$a_{1,2} = \frac{B_1}{2} \pm \sqrt{\left(\frac{B_1}{2}\right)^2 + 2 - B_2}. \quad (13)$$

Thus, the dispersion equation (12) has two solutions; physically, this means that there are two independent spectra of natural waves in the structure considered here, each being characterized by an individual dispersion equation and an individual Bloch wave number. These equations are obtained from Eq. (12) upon substituting (13) into it and replacing ξ with $\exp(i\bar{k}_{1,2}d)$. As a result, we arrive at

$$\cos\bar{k}_1 d = -\frac{a_1}{2}, \quad (14)$$

$$\cos\bar{k}_2 d = -\frac{a_2}{2}. \quad (15)$$

They describe two band spectra of natural waves in the structure, the transmission bands in these spectra being specified by the conditions $|a_1| \le 2$ and $|a_2| \le 2$, where a_1 and a_2 are real-valued numbers. The respective values of $\bar{k}_{1,2}$ are real. In all other cases, there is no transmission, and $\bar{k}_{1,2}$ are complex-valued numbers [13]. Since Eqs. (14) and (15) are independent of each other, the transmission bands of these spectra may overlap or may be superimposed. We note that the independence of the dispersion equations for the Bloch wave numbers $\bar{k}_1$ and $\bar{k}_2$ is due to the disregard of dissipation processes in the problem under study.

BAND STRUCTURE OF THE SPECTRUM

The results obtained by numerically solving the problem being considered are represented in the form of two independent band spectra of natural waves in a periodic structure at a fixed induction of the applied magnetic field. For our calculations, we took an *n*-InSb semiconductor layer characterized by an electron concentration of 3×10^{14} cm^{-3} and by the values of $\varepsilon_L = 17.8$ and $d_1 = 0.01$ cm and a dielectric layer characterized by the values of $\varepsilon_2 = 2.0$ and $d_2 = 0.03$ cm.

Preliminarily, we would like to note the following:

1. In order to obtain deeper insight into the process of formation of the band spectrum of the structure, it is necessary to take into account the frequency dependence of the elements of the tensor $\hat{\varepsilon}$ and of the Voigt dielectric permittivity ε_V. For the magnetic-field-induction value of $B_0 = 0.05$ T, such dependences are displayed in Fig. 2, where one can single out some characteristic regions. These are regions to the left and to the right of the cyclotron-resonance frequency, at which the dielectric-permittivity-tensor components ε_1 and ε_2 diverge, changing sign. Moreover, there is a characteristic point at the plasma frequency, where $\varepsilon_3 = 0$; in addition, the Voigt dielectric permittivity ε_V goes to infinity at the hybrid-cyclotron-resonance frequency $\omega_h = \sqrt{\omega_H^2 + \omega_P^2}$. In these regions, there must arise singularities in the band structure of the spectrum as well.

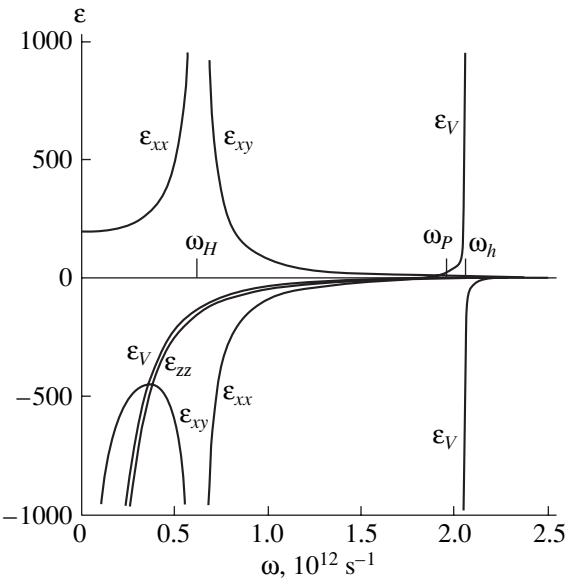


Fig. 2. Frequency dependence of the elements of the tensor $\hat{\varepsilon}$ and of the Voigt dielectric permittivity ε_V for a semiconductor layer of the periodic structure under consideration.

2. Expressions (3) and (6) for $k_{z1,2}^2$ and k_z^2, respectively, depend on the wave-vector components k_x and k_y only through the combination $k_{xy}^2 = k_x^2 + k_y^2$. Therefore, one can assume that the spectrum will also be determined by the quantity k_{xy} exclusively. Indeed, numerical calculations reveal that, under the condition $k_{xy} = \text{const}$, identical values of $\omega = f(k_x, k_y)$ are obtained for different values of the ratio k_y/k_x. From the physical point of view, this means that the rotation of the plane spanned by the *z* axis and the vector $\mathbf{k_{xy}}$—this is the plane in which the propagation of waves occurs along the *z* axis (see Fig. 1)—does not change the pattern of the spectrum. It follows that the use of the quantity k_{xy} (more precisely, $k_{xy}d$) for a variable parameter is quite reasonable.

3. In the periodic structure being considered, four partial waves—those that have the transverse wave numbers k_{z1} and k_{z2} in a semiconductor layer and those

that have identical transverse wave numbers k_z but different (orthogonal) polarizations in a dielectric layer—are involved in the process of formation of the band spectrum of natural waves. We note that, as follows from expressions (3) and (6), k_{z1} and k_{z2} may be real-valued, pure imaginary, or complex-valued quantities, while k_z may be a real-valued or pure imaginary quantity. Various combinations of these possibilities for these three transverse wave numbers determine the possible existence of one type of natural wave in a given section of the band spectrum or another, as well as the distribution of electric and magnetic fields in the layers of the structure. In analyzing the band spectra, it is therefore of importance to know positions of the boundaries between regions where the transverse wave numbers are of different complex character. By using Eq. (3), we can show that, in a semiconductor layer, the boundary between real and pure imaginary values of $k_{z1,2}$ are determined by the relations

$$k_{xy} = k_0\sqrt{\varepsilon_3} \tag{16}$$

(for frequencies above ω_P) and

$$k_{xy} = k_0\sqrt{\varepsilon_V}. \tag{17}$$

The boundaries between complex and real or pure imaginary values of $k_{z1,2}$ are specified by the relations

$$k_{xy} = \frac{k_0}{\omega_H}\sqrt{2\varepsilon_L(\omega^2 - \omega_P^2)\left[1 \pm \sqrt{1 - \left(\frac{\omega_H}{\omega}\right)^2}\right]}. \tag{18}$$

They occur in the region of frequencies for which $\omega > \omega_P, \omega_H$. For $\omega \gg \omega_H$, relation (18) for one of the transverse wave numbers (with a minus sign in front of the square root) reduces to (16).

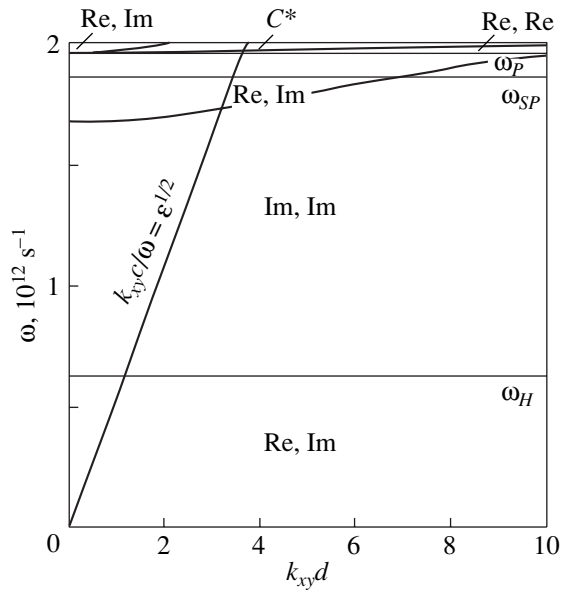


Fig. 3. Regions where the transverse wave numbers k_{z1} and k_{z2} in the semiconductor layers of the structure are of different complex character (C^* stands for complex conjugate quantities).

For the structure under study, the boundaries between the regions where k_z is of different complex character are shown in Fig. 3 at $B_0 = 0.05$ T. For a semiconductor, these regions are bounded by the displayed curves and the dashed lines corresponding to the frequencies ω_P, ω_{SP}, and ω_H. For a dielectric layer, the light-propagation line $k_{xy}c/\omega = \varepsilon^{1/2}$ appears as the respective boundary, with k_z values being real to the left of it and pure imaginary to the right.

The total spectrum of our periodic structure is shown in Fig. 4a according to calculations at the magnetic-field-induction value of $B_0 = 0.05$ T and $k_y/k_x = 2$. This spectrum was obtained by taking the sum of two independent spectra corresponding to the Bloch wave numbers $\bar{k}_1$ and $\bar{k}_2$ that are determined by the dispersion equations (14) and (15) and which are shown in Figs. 4b and 4c. The spectrum consists of alternating transmission and nontransmission bands whose width is a function of $k_{xy}d$. In the structure of the spectrum, there are several specific regions characterized by the corresponding types of natural waves in our periodic structure. We will now consider some of them.

Region of collective surface waves. It includes two transmission bands occurring in the region of frequencies below the plasma frequency and also existing in the absence of a magnetic field. This property of a periodic medium—that is, the property of being transparent to electromagnetic waves at frequencies where the semiconductor layers entering into its composition are not transparent—was indicated in [1]. Such waves can be classified as unusual waves. A feature that is peculiar to the present case and which distinguishes it from the case considered in [1] is that, now, four partial waves (rather than two) determining the topology of the spectrum (in this region inclusive) are involved in the formation of the bands. This leads both to a change in the shapes of transmission bands and to a change in the character of the field distributions over the layers.

At large $k_{xy}d$, the width of these transmission bands decreases; they approach a horizontal asymptote corresponding to the surface-plasmon frequency $\omega_{SP} = \sqrt{\varepsilon_L\omega_P^2/(\varepsilon_L + \varepsilon_2)}$. In this region of frequencies, the transverse wave numbers are pure imaginary (see Fig. 3) and correspond to partial waves whose fields decrease from the boundaries of each layer according to an exponential law. We also note that, at the uppermost boundary of this transmission band, there exists a segment of negative group velocity. The shape of the transmission bands depends on the induction of the applied magnetic field.

Region of waveguide plasma waves. It is represented by a narrow transmission band having an asymptote at the plasma frequency. This region does not exist in the absence of a magnetic field. For partial waves in

a semiconductor layer, the transverse wave numbers are real-valued (see Fig. 3); that is, these are waves of the waveguide type in each layer. In a dielectric layer, the wave numbers are pure imaginary, while the fields are of a superficial character; that is, their amplitudes decrease from the boundaries of dielectric layers according to an exponential law.

Region of cyclotron waves. The region consists of a number of bands lying below the cyclotron frequency; the bandwidth decreases as one approaches the cyclotron frequency, tending to zero in the limit. This is because, in the vicinity of this frequency, the wave numbers k_{z1} and k_{z2} can take indefinitely large values in view of the divergence of the components of the tensor $\hat{\varepsilon}$. Since trigonometric functions of the arguments $k_{z1}d_1$ and $k_{z2}d_1$ appear in the dispersion equation, this leads to the emergence of an infinite number of transmission and nontransmission bands. In [6], a similar effect was found in the case of a "transverse" propagation of waves at frequencies below the hybrid-resonance frequency. It goes without saying that such a fragmentation of bands is possible only if dissipative processes are disregarded. In the presence of dissipation, the width of transmission and nontransmission bands on the frequency axis cannot be less than the dissipation frequency.

In the frequency region $\omega \le \omega_H$ being considered, the expressions for the transverse wave numbers in the layers under the condition $k_{xy} \gg k_0$ assume the form

$$k_{z1} \approx \pm \sqrt{\left|\frac{\varepsilon_1}{\varepsilon_3}\right|} k_{xy}, \tag{19}$$

$$k_{z2} \approx k_z \approx \pm i k_{xy}. \tag{20}$$

Here, we have taken into account the frequency dependence in the components of the tensor $\hat{\varepsilon}$ of the semiconductor dielectric permittivity (see Fig. 2). From the above relations, it follows that, in the approximation used here, one partial wave in a semiconductor layer, that of wave number k_{z1}, has a waveguide character, while the other, that of wave number k_{z2}, has a surface character. This situation, in which volume waves must be "admixed" to surface waves in order to satisfy the boundary conditions, was classified in [14] as that in which there emerge pseudosurface modes.

In the upper part of the band spectrum, the transverse wave number in the dielectric material to the left of the light-propagation line is real-valued; in the same region, one transverse wave number is real-valued, while the other is pure imaginary (see Fig. 3), that is, this is also the region of pseudosurface modes. Here, there exist a few broad transmission bands corresponding to Bloch wave numbers $\bar{k}_{1,2}$. As can be seen from

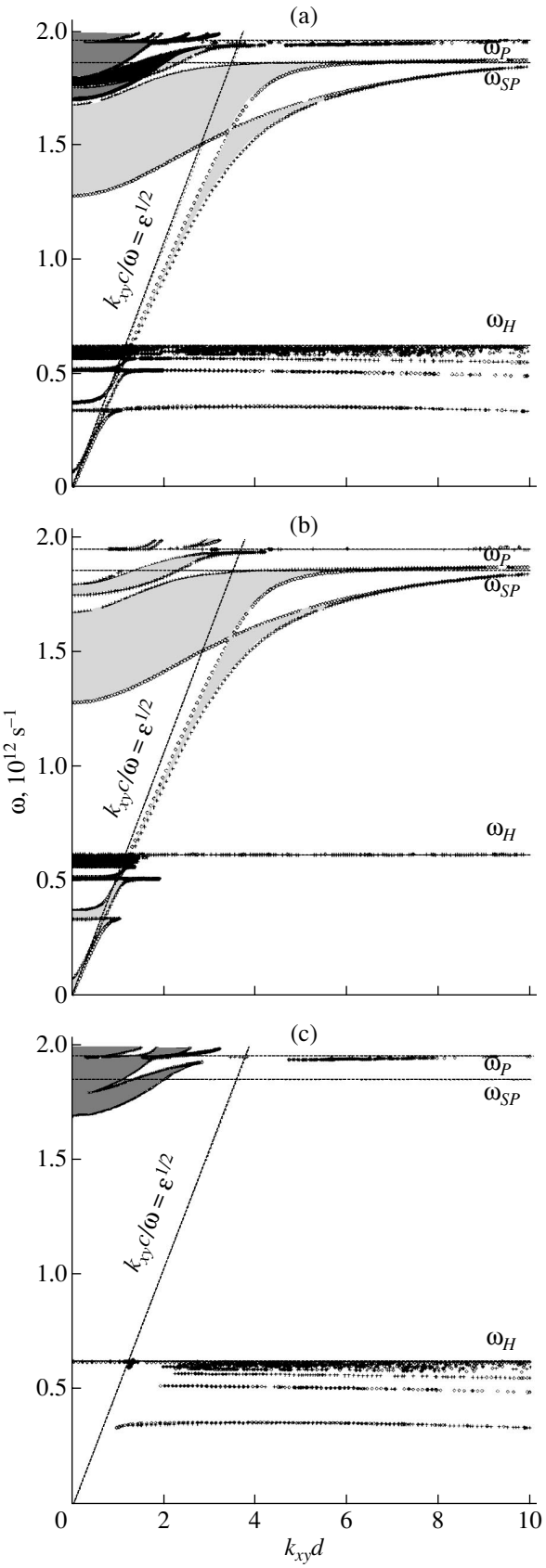


Fig. 4. Band spectra of a periodic structure: (a) total spectrum, (b) spectrum for the Bloch wave number $\bar{k}_1$, and (c) spectrum for the Bloch wave number $\bar{k}_2$.

Fig. 4, these bands partly overlap, forming a transmission band common to $\bar{k}_1$ and $\bar{k}_2$.

CONCLUSION

The band spectrum of a periodic layered semiconductor structure placed in a magnetic field aligned with the periodicity axis has been calculated for the case of wave propagation at some angle with respect to the magnetic field. A method for deriving the dispersion equation is given. It has been shown numerically that, for the periodic-structure configuration being considered, the left-hand side of the dispersion equation can be represented as the product of two polynomials of second degree, and an analysis of the properties of such an equation has been performed.

It has been established that, in this structure, there are two independent spectra of natural waves in the absence of dissipation. It has been shown that the band spectrum consists of collective plasma waves and numerous bands of cyclotron waves.

The regions of the existence of natural waves of this spectrum have been classified. It has been found that the transmission bands of the two independent spectra of the periodic structure in question can be superimposed and supplement one another.

APPENDIX 1

The coefficients C_i and D_i in expressions (4) for the field components in a semiconductor are given by

$$C_{1,2} = \left\{ ik_y\left[\varepsilon_1\left(k_0^2 - \frac{k_{xy}^2}{\varepsilon_3}\right) - k_{z1,z2}^2\right] - \varepsilon_2 k_x\left(k_0^2 - \frac{k_{xy}^2}{\varepsilon_3}\right)\right\} e^{i(k_x x + k_y y)},$$

$$C_{3,4} = \left\{ -ik_x\left[\varepsilon_1\left(k_0^2 - \frac{k_{xy}^2}{\varepsilon_3}\right) - k_{z1,z2}^2\right] - \varepsilon_2 k_y\left(k_0^2 - \frac{k_{xy}^2}{\varepsilon_3}\right)\right\} e^{i(k_x x + k_y y)},$$

$$D_{1,2} = k_{z1,z2}\left\{ \frac{k_y}{k_0}\left[\varepsilon_1\left(k_0^2 - \frac{k_{xy}^2}{\varepsilon_3}\right) - k_{z1,z2}^2\right] - i\varepsilon_2 k_0 k_y \right\} e^{i(k_x x + k_y y)},$$

$$D_{3,4} = k_{z1,z2}\left\{ \frac{k_y}{k_0}\left[\varepsilon_1\left(k_0^2 - \frac{k_{xy}^2}{\varepsilon_3}\right) - k_{z1,z2}^2\right] + i\varepsilon_2 k_0 k_x \right\} e^{i(k_x x + k_y y)}$$

For the elements of the matrix $\hat{S}$ for a semiconductor, we have

$$S_{11} = C_5(C_1C_4\cos\alpha - C_2C_3\cos\beta),$$

$$S_{21} = C_3C_4C_5(\cos\alpha - \cos\beta),$$

$$S_{12} = -C_1C_2C_5(\cos\alpha - \cos\beta),$$

$$S_{22} = -C_5(C_2C_3\cos\alpha - C_1C_4\cos\beta),$$

$$S_{13} = D_5(C_1D_4\sin\alpha - C_2D_3\sin\beta),$$

$$S_{23} = D_5(C_3D_4\sin\alpha - C_4D_3\sin\beta),$$

$$S_{14} = -D_5(C_1D_2\sin\alpha - C_2D_1\sin\beta),$$

$$S_{24} = -D_5(C_3D_2\sin\alpha - C_4D_1\sin\beta),$$

$$S_{31} = -C_5(D_1C_4\sin\alpha - D_2C_3\sin\beta),$$

$$S_{41} = -C_5(D_3C_4\sin\alpha - D_4C_3\sin\beta),$$

$$S_{32} = C_5(D_1C_2\sin\alpha - D_2C_1\sin\beta),$$

$$S_{42} = C_5(D_3C_2\sin\alpha - D_4C_1\sin\beta),$$

$$S_{33} = D_5(D_1C_4\cos\alpha - D_2D_3\cos\beta),$$

$$S_{43} = D_3D_4D_5(\cos\alpha - \cos\beta),$$

$$S_{34} = -D_1D_2D_5(\cos\alpha - \cos\beta),$$

$$S_{44} = -D_5(D_2D_3\cos\alpha - D_1D_4),$$

where $\alpha = k_1 z$, $\beta = k_{z2} z$, $C_5 = (C_1C_4 - C_2C_3)^{-1}$, and $D_5 = (D_1D_4 - D_2D_3)^{-1}$.

APPENDIX 2

The coefficients F_i in expressions (7) for the field components in a dielectric material are given by

$$F_1 = \cos k_z z, \quad F_2 = i\frac{k_x k_y}{k_0 k_z}\sin k_z z,$$

$$F_3 = i\frac{(k_x k_y)^2 + \varepsilon(k_0 k_z)^2}{k_0 k_z(k_y^2 - \varepsilon k_0^2)}\sin k_z z,$$

$$F_4 = i\frac{k_y^2 - \varepsilon k_0^2}{k_0 k_z}\sin k_z z.$$

The matrix $\hat{D}$ for a dielectric material has the form

$$\hat{D} = \begin{pmatrix} F_1 & 0 & F_2/\varepsilon & -F_3/\varepsilon \\ 0 & F_1 & F_4/\varepsilon & -F_2/\varepsilon \\ -F_2 & F_3 & F_1 & 0 \\ -F_4 & F_2 & 0 & F_1 \end{pmatrix}.$$

APPENDIX 3

The coefficients in the dispersion equation that are expressed in terms of the transformation matrix $\hat{P}$ for one period of the structure are given by

$$B_1 = -\sum_{i=1}^{2}\sum_{j=i+1}^{3}\sum_{k=j+1}^{4}(p_{ii}p_{jj}p_{kk} + p_{ij}p_{jk}p_{ki} + p_{ik}p_{ji}p_{kj} - p_{ii}p_{jk}p_{kj} - p_{jj}p_{ki}p_{ik} - p_{kk}p_{ij}p_{ji}),$$

$$B_2 = \sum_{i=1}^{3}\sum_{j=i+1}^{4}(p_{ii}p_{jj} - p_{ij}p_{ji}),$$

$$B_3 = -\sum_{i=1}^{4}p_{ii}.$$

REFERENCES

1. F. G. Bass, A. A. Bulgakov, and A. P. Tetervov, *High-Frequency Properties of Semiconductors with Superlattices* (Nauka, Moscow, 1989).
2. A. Yariv and P. Yeh, *Optical Waves in Crystals: Propagation and Control of Laser Radiation* (Wiley, New York, 1984; Mir, Moscow, 1987).
3. A. C. Baynham and A. D. Boardman, J. Phys. C **2**, 619 (1969).
4. V. L. Ginzburg, *The Propagation of Electromagnetic Waves in Plasmas* (Nauka, Moscow, 1967; Pergamon, Oxford, 1970).
5. R. S. Brazis, A. S. Mironas, and Yu. K. Pozhela, Litovsk. Fiz. Sb. **14** (1), 95 (1974).
6. A. A. Bulgakov and O. V. Shramkova, Radiotekh. Élektron. (Moscow) **46**, 236 (2001).
7. J. J. Brion, R. F. Wallis, A. Hardstein, and E. Burstein, Phys. Rev. Lett. **28**, 1455 (1972).
8. R. F. Wallis, J. J. Brion, E. Burstein, and A. Hartstein, Phys. Rev. B **9**, 3424 (1974).
9. Yu. V. Bludov, Dop. NAN Ukr., No. 12, 79 (1999).
10. A. I. Akhiezer, I. A. Akhiezer, R. V. Polovin, *et al.*, *Plasma Electrodynamics*, Ed. by A. I. Akhiezer (Nauka, Moscow, 1974; Pergamon, Oxford, 1975).
11. A. G. Gurevich, *Ferrites at Microwaves* (Fizmatgiz, Moscow, 1960).
12. M. Born and É. Wolf, *Principles of Optics* (Pergamon, Oxford, 1969; Nauka, Moscow, 1973).
13. S. Yu. Karpov and S. N. Stolyarov, Usp. Fiz. Nauk **163** (1), 63 (1993) [Phys. Usp. **36**, 1 (1993)].
14. R. S. Brazis, Litovsk. Fiz. Sb. **21** (4), 73 (1981).

Translated by A. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1380–1388. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 22–30.

GASES AND LIQUIDS

On the Stability of Nonaxisymmetric Modes of a Spatially Charged Jet of a Viscous Dielectric Liquid

S. O. Shiryaeva, A. I. Grigor'ev, and T. V. Levchuk

Demidov Yaroslavl State University, Sovetskaya ul. 14, Yaroslavl, 150000 Russia

e-mail: shir@uniyar.ac.ru

Received April 24, 2003

Abstract—A dispersion equation is derived for nonaxisymmetric capillary vibrations of a spatially charged jet of a viscous dielectric liquid. It is shown that the critical conditions with respect to the instability of nonaxisymmetric jet modes are less stringent than those for axisymmetric modes, this effect being more pronounced for liquids characterized by lower values of the dielectric permittivity. *© 2003 MAIK "Nauka/Interperiodica".*

1. Investigation of capillary vibrations in charged jets of an electrically conducting or a dielectric liquid and of its stability and conditions of its disintegration into drops is of interest in connection with numerous application of these phenomena in various fields of technical physics and technology (see [1, 2] and references therein). In view of the aforesaid, the physical object in question has been repeatedly studied experimentally and theoretically, both within a linear and within a nonlinear formulation of the problem [1–12]. Despite this, some questions concerning the stability of jets have yet to be clarified conclusively. In particular, this is so for an analysis of stability of nonaxisymmetric vibrational modes and is due to the fact that the majority of the previous investigations were aimed, in order to meet requirements of practice, at obtaining fluxes of monodispersed drops [2]. Nonetheless, one has to consider the disintegration of nonaxisymmetric jets into drops in exploring the spontaneous fragmentation of jets ejected from the vertices of freely falling drops [13] and from meniscuses of liquids at the end faces of capillaries [14, 15] and in dealing with the instabilities of jets with respect to their surface charge. General regularities of the loss of stability were analyzed in [12] for nonaxisymmetric modes of superficially charged jets formed by ideally conducting liquids, and it was found there that the critical conditions for the realization of the instability of nonaxisymmetric modes are less stringent than those for axisymmetric modes, this effect becoming more pronounced with increasing viscosity. In this connection, it is of interest to examine stability of nonaxisymmetric modes for another limiting case, that of spatially charged jets of ideally conducting liquids. This situation is of importance in connection with creating electrically controlled jets of charged liquid-hydrogen drops for laser-induced thermonuclear fusion [16], accelerators of macroscopic particles [17, 18], and neutrino detectors [19].

2. Suppose that we have an infinite cylindrical jet formed by a viscous incompressible liquid of mass density ρ, kinematic viscosity ν, dielectric permittivity ε, and the coefficient of surface tension σ and that the jet, which has a radius R, moves at a constant velocity $\mathbf{U}_0$ along the symmetry axis. Within the frozen-charge model, we assume that the charge is distributed uniformly and denote its density by μ. Since we consider an infinite jet, we go over, in order to simplify the problem, to inertial reference frame comoving with the jet. It is obvious that, in this reference frame, the velocity field $\mathbf{U}(\mathbf{r}, t)$ for the flow of the liquid in the jet is completely determined by possible capillary vibrations (of thermal origin, for example) of the jet surface and is of the same order of smallness as the amplitude of vibrations. We will seek critical conditions for the realization of the instability of capillary vibrations of the drop surface.

All the ensuing calculations will be performed in the system of cylindrical coordinates with the z axis aligned with the jet symmetry axis, the respective unit vector $\mathbf{n}_z$ being directed along the velocity vector $\mathbf{U}_0$. We represent the equation of the jet surface perturbed by a capillary wave motion in the form

$$r = R + \xi(z, \phi, t), \quad |\xi| \ll R.$$

The mathematical formulation of the problem of calculating capillary vibrations of a jet includes the equations of fluid dynamics and electrostatics (under the assumption that the velocity of the motion of the liquid in the jet is much less than relativistic velocities),

$$\frac{d\mathbf{U}}{dt} = -\frac{1}{\rho}\nabla P + \nu\Delta\mathbf{U}; \quad \nabla\cdot\mathbf{U} = 0;$$

$$\Delta\Phi^{\text{in}} = -4\pi\frac{\mu}{\varepsilon}; \quad \Delta\Phi^{\text{ex}} = 0;$$

1063-7842/03/4811-1380$24.00

the conditions of boundedness,

$$r \longrightarrow 0\colon \Phi^{in} \longrightarrow \text{const}; \quad |\mathbf{U}| = \text{const} < \infty;$$

$$r \longrightarrow \infty\colon \nabla\Phi^{ex} \longrightarrow 0;$$

and the boundary conditions at the free jet surface

$$r = R + \xi\colon \Phi^{in} = \Phi^{ex}, \quad \varepsilon\frac{\partial\Phi^{in}}{\partial n} = \frac{\partial\Phi^{ex}}{\partial n};$$

$$-\frac{\partial\xi}{\partial t} + \mathbf{U}\cdot\nabla[r - (R + \xi(z, \phi, t))] = 0;$$

$$\frac{\partial U_z}{\partial r} + \frac{\partial U_r}{\partial z} = 0;$$

$$\frac{\partial U_\phi}{\partial r} + \frac{1}{r}\frac{\partial U_r}{\partial\phi} - \frac{1}{r}U_\phi = 0;$$

$$-P(\mathbf{r}, t) - P_0 + 2\nu\frac{\partial U_r}{\partial r} - P_e + P_\sigma = 0.$$

Here, U_r, U_ϕ, and U_z are the components of the velocity field in the system of cylindrical coordinates; $P(\mathbf{r}, t)$ is a hydrodynamic pressure; P_0 is atmospheric pressure; P_e is the electric-field pressure (see Appendix); P_σ is the pressure of the surface-tension forces ($P_\sigma = \sigma\,\mathrm{div}\,\mathbf{n}$); $\mathbf{n}$ is a unit vector orthogonal to the jet surface; and Φ^{in} and Φ^{ex} are the electrostatic potentials within and beyond the jet, respectively.

The solution to the above problem will be sought in the form of an expansion in the small parameter $|\xi|/R$.

In the zeroth approximation, we obtain an immobile column of a liquid and the well-known expression for the electric field pressure at the surface of a uniformly charged infinite cylinder with a fixed radius.

3. Going over to a system of dimensional variables where $R = 1$, $\rho = 1$, and $\sigma = 1$ and preserving the former notation for all physical quantities, we write the problem under consideration in the linear approximation in $|\xi|$ as

$$r = 1 + \xi(\phi, z, t); \quad |\xi| \ll 1; \tag{1}$$

$$\frac{\partial\mathbf{U}}{\partial t} = -\nabla P + \nu\Delta\mathbf{U}; \tag{2}$$

$$\nabla\mathbf{U} = 0; \tag{3}$$

$$\Delta\varphi^{in} = 0; \quad \Delta\varphi^{ex} = 0; \tag{4}$$

$$r \longrightarrow 0\colon \varphi^{in} \longrightarrow 0, \quad |\mathbf{U}| < \infty, \tag{5}$$

$$r \longrightarrow \infty\colon \varphi^{ex} \longrightarrow 0; \tag{6}$$

$$r = 1\colon \varphi^{in} + \frac{\partial\Phi_0^{in}}{\partial r}\xi = \varphi^{ex} + \frac{\partial\Phi_0^{ex}}{\partial r}\xi; \tag{7}$$

$$\varepsilon\left(\frac{\partial\varphi^{in}}{\partial r} + \frac{\partial^2\Phi_0^{in}}{\partial r^2}\xi\right) = \frac{\partial\varphi^{ex}}{\partial r} + \frac{\partial^2\Phi_0^{ex}}{\partial r^2}\xi; \tag{8}$$

$$-\frac{\partial\xi}{\partial t} + U_r = 0; \tag{9}$$

$$\frac{\partial U_\phi}{\partial r} + \frac{\partial U_r}{\partial\phi} - U_\phi = 0; \tag{10}$$

$$\frac{\partial U_z}{\partial r} + \frac{\partial U_r}{\partial z} = 0; \tag{11}$$

$$-p + 2\nu\frac{\partial U_r}{\partial r} - p_e + p_\sigma = 0. \tag{12}$$

In Eqs. (1)–(12), φ^{in}, φ^{ex}, p, p_e, and p_σ are corrections to, respectively, the electric potential within the liquid, the electric potential beyond the liquid, the hydrodynamic pressure, the pressure of electric forces, and the pressure of surface-tension forces from capillary vibrations of the jet surface.

Expanding the Laplace pressure and the electric-field pressure at the jet surface in the small quantity $|\xi|$, we can easily obtain expressions of the first order of smallness for the quantities p_e and p_σ. Specifically, they are given by (see Appendix; see also [12])

$$p_\sigma = -\left(\xi + \frac{\partial^2\xi}{\partial\phi^2} + \frac{\partial^2\xi}{\partial z^2}\right); \tag{13}$$

$$p_e = 2\pi\mu^2\left(1 - \frac{\varepsilon - 1}{2\varepsilon}\right)\xi - \mu\left(\varphi^{ex} + \frac{\varepsilon - 1}{2\varepsilon}\frac{\partial\varphi^{ex}}{\partial r}\right). \tag{14}$$

4. We will solve the set of Eqs. (2) and (3) by the method of operator scalarization (for details, see [20]; see also [12]), expanding the velocity field $\mathbf{U}(\mathbf{r}, t)$ in the sum of three orthogonal vector fields with the aid of differential vector operators $\mathbf{N}_i$,

$$\mathbf{U}(\mathbf{r}, t) = \sum_{i=1}^{3}\hat{\mathbf{N}}_i\psi_i(\mathbf{r}, t) \quad (i = 1, 2, 3), \tag{15}$$

that satisfy the orthogonality conditions

$$\hat{\mathbf{N}}_j^+\hat{\mathbf{N}}_i = 0 \quad (\text{for } i \neq j; \quad i, j = 1, 2, 3) \tag{16}$$

and commute with the Laplace operator:

$$\Delta\hat{\mathbf{N}}_i = \hat{\mathbf{N}}_i\Delta. \tag{17}$$

In expressions (15) and (16), $\psi_i(\mathbf{r}, t)$ are unknown scalar functions and $\hat{\mathbf{N}}_j^+$ are the Hermitian conjugates of $\hat{\mathbf{N}}_j$.

Since the equilibrium shape of the jet is axisymmetric, it is convenient to take the operators $\hat{\mathbf{N}}_i$ in the form

$$\hat{\mathbf{N}}_1 = \nabla; \quad \hat{\mathbf{N}}_2 = \nabla\times\mathbf{e}_z; \quad \hat{\mathbf{N}}_3 = \nabla\times(\nabla\times\mathbf{e}_z).$$

From (15), it follows that, in the system of cylindrical coordinates, the expressions for the components of the velocity field $\mathbf{U}(\mathbf{r}, t)$ in terms of the scalar function $\psi_i(\mathbf{r}, t)$ are given by

$$U_r = \frac{\partial\psi_1}{\partial r} + \frac{1}{r}\frac{\partial\psi_2}{\partial\phi} + \frac{\partial^2\psi_3}{\partial z\partial r};$$

$$U_\phi = \frac{1}{r}\frac{\partial\psi_1}{\partial\phi} - \frac{\partial\psi_2}{\partial r} + \frac{1}{r}\frac{\partial^2\psi_3}{\partial z\partial\phi}; \quad (18)$$

$$U_z = \frac{\partial\psi_1}{\partial z}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\psi_3}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\psi_3}{\partial\phi^2}\right].$$

Substituting expressions (15) and (18) into the set of Eqs. (2) and (3) and using the operator properties (16) and (17), we obtain the set of scalar equations

$$\Delta\psi_1 = 0; \quad \Delta\psi_i - \frac{1}{\nu}\frac{\partial\psi_i}{\partial t} = 0 \ (i = 2, 3); \quad (19)$$

$$p = -\frac{\partial\psi_1}{\partial t}. \quad (20)$$

With the aid of relations (13), (14), (18), and (20), we transform the boundary conditions (9)–(12) into the following boundary conditions for the unknown functions ψ_i and ε:

$$r = 1: \frac{\partial\xi}{\partial t} - \left[\frac{\partial\psi_1}{\partial r} + \frac{1}{r}\frac{\partial\psi_2}{\partial\phi} + \frac{\partial^2\psi_3}{\partial z\partial r}\right] = 0;$$

$$2\frac{\partial}{\partial\phi}\left\{\frac{\partial\psi_1}{\partial r} - \psi_1\right\} - \left\{\frac{\partial^2\psi_2}{\partial r^2} - \frac{\partial\psi_2}{\partial r} - \frac{\partial^2\psi_2}{\partial\phi^2}\right\}$$

$$+ 2\frac{\partial^2}{\partial z\partial\phi}\left\{\frac{\partial\psi_3}{\partial r} - \psi_3\right\} = 0;$$

$$\frac{\partial}{\partial r}\left\{\frac{\partial\psi_1}{\partial z} - \left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\psi_3}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\psi_3}{\partial\phi^2}\right]\right\} \quad (21)$$

$$+ \frac{\partial}{\partial z}\left\{\frac{\partial\psi_1}{\partial r} + \frac{1}{r}\frac{\partial\psi_2}{\partial\phi} + \frac{\partial^2\psi_3}{\partial z\partial r}\right\} = 0;$$

$$\frac{\partial\psi_1}{\partial t} + 2\nu\frac{\partial}{\partial r}\left\{\frac{\partial\psi_1}{\partial r} + \frac{1}{r}\frac{\partial\psi_2}{\partial\phi} + \frac{\partial^2\psi_3}{\partial z\partial r}\right\}$$

$$- 2\pi\mu^2\left(1 - \frac{\varepsilon - 1}{2\varepsilon}\right)\xi + \mu\left(\varphi^{ex} + \frac{\varepsilon - 1}{2\varepsilon}\frac{\partial\varphi^{ex}}{\partial r}\right)$$

$$- \left(\xi + \frac{\partial^2\xi}{\partial\phi^2} + \frac{\partial^2\xi}{\partial z^2}\right) = 0.$$

Since the functions ξ, φ^{ex}, φ^{in}, and ψ_i describe small deviations from the equilibrium state, we assume, in order to analyze its stability, that the time dependence of all small quantities has the exponential form

$$\xi, \varphi^{ex}, \varphi^{in}, \psi_i \sim \exp(st),$$

where s is the complex-values frequency.

For Eqs. (19) and (4), solutions satisfying the boundedness conditions (5) and (6) will then be sought in the system of cylindrical coordinates in the form of expansions in waves running over the unperturbed cylindrical surface of the jet along the z axis:

$$\psi_1 = \int_0^\infty \sum_{m=0}^\infty C_1 I_m(kr)\exp(im\phi)\exp(ikz)\exp(st)dk;$$

$$\psi_i = \int_0^\infty \sum_{m=0}^\infty C_i I_m(lr)\exp(im\phi)\exp(ikz)\exp(st)dk \quad (22)$$

$$(i = 2, 3);$$

$$\varphi^{ex} = \int_0^\infty \sum_{m=0}^\infty C_4 K_m(kr)\exp(im\phi)\exp(ikz)\exp(st)dk; \quad (23)$$

$$\varphi^{in} = \int_0^\infty \sum_{m=0}^\infty C_5 I_m(kr)\exp(im\phi)\exp(ikz)\exp(st)dk.$$

The function $\xi(z, \phi, t)$ will be represented in the form of the analogous expansion

$$\xi(z, \phi, t) = \int_0^\infty \sum_{m=0}^\infty D\exp(im\phi)\exp(ikz)\exp(st)dk. \quad (24)$$

In expressions (22)–(24), k is the wave number; $l^2 \equiv k^2 + s/\nu$; m is the azimuthal number characterizing the deviation of the solutions from axial symmetry; $I_m(x)$ and $K_m(x)$ are modified Bessel functions of the first and the second kind, respectively; and C_i (where $i = 1, 2, 3, 4, 5$) and D are expansion coefficients, which depend on m and k.

Using the solutions given by (23) and (24) and considering that the functions $\exp(im\phi)$ and $\exp(ikz)$ are orthonormalized,

$$\int_0^{2\pi} \exp[i(m_1 - m_2)\phi]d\phi = \delta_{m_1, m_2};$$

$$\int_{-\infty}^{\infty} \exp[i(k_1 - k_2)z]dz = \delta(k_1 - k_2), \quad (25)$$

where δ_{m_1, m_2} is the Kronecker delta symbol and $\delta(k_1 - k_2)$ is a Dirac delta function, we can easily derive, from the boundary conditions (7) and (8), the relations

between the coefficients C_4, C_5, and D. Specifically, we have

$$C_4 = \frac{2\pi\mu D}{K_m(k)}\frac{2+(\varepsilon-1)kh(k)}{\varepsilon k[h(k)-g(k)]};$$

$$C_5 = \frac{2\pi\mu D}{I_m(k)}\frac{2\varepsilon+(\varepsilon-1)kg(k)}{\varepsilon^2 k[h(k)-g(k)]}; \qquad (26)$$

$$h(k) \equiv I'_m(k)/I_m(k); \quad g(k) \equiv K'_m(k)/K_m(k).$$

Substituting solutions (22) and (23) with allowance for (26) and expansion (24) into the boundary conditions (21) and using relations (25), we arrive at the following set of equations for the unknown coefficients D and C_i (i = 1, 2, 3):

$$Ds - C_1 kI'_m(k) - C_2 imI_m(l) - C_3 iklI'_m(l) = 0;$$

$$C_1 2im(kI'_m(k) - I_m(k)) + C_2(lI'_m(l) - m^2 I_m(l)$$
$$- l^2 I''_m(l)) + C_3 2mk(I_m(l) - lI'_m(l)) = 0;$$

$$C_1 2ik^2 I'_m(k) - C_2 mkI_m(l) - C_3(l^3 I'''_m(l) + l^2 I''_m(l)$$
$$+ l(k^2 - m^2 - 1)I'_m(l) + 2m^2 I_m(l)) = 0; \qquad (27)$$

$$D\left\{k^2 + m^2 - 1 - 2\pi\mu^2\left[1 - \frac{\varepsilon-1}{2\varepsilon}\right.\right.$$

$$\left.\left. - \frac{(2+(\varepsilon-1)kh(k))(2\varepsilon+(\varepsilon-1)kg(k))}{2\varepsilon^2 k[h(k)-g(k)]}\right]\right\}$$

$$+ C_1(sI_m(k) + 2\nu k^2 I''_m(k)) + C_2 2\nu im(lI'_m(l) - I_m(l))$$
$$+ C_3 2\nu ikl^2 I''_m(l) = 0.$$

Here, primes denote the derivatives of Bessel functions of order m with respect to the argument; these derivatives can be expressed in terms of Bessel functions of order m and (m + 1) with the aid of the recursion relations

$$I'_m(x) = I_{m+1}(x) + \frac{m}{x}I_m(x) = I_{m-1}(x) - \frac{m}{x}I_m(x);$$

$$I''_m(x) = -\frac{1}{x}I_{m+1}(x) + \left(1 + \frac{m(m-1)}{x^2}\right)I_m(x);$$

$$I'''_m(x) = \left(1 + \frac{m^2+2}{x^2}\right)I_{m+1}(x)$$
$$+ \frac{m-1}{x}\left(1 + \frac{m(m-2)}{x^2}\right)I_m(x);$$

$$K'_m(x) = \frac{m}{x}K_m(x) - K_{m+1}(x).$$

We recall that the set of homogeneous equations (27) has a nontrivial solution only if its determinant is equal to zero, $\det[a_{ij}] = 0$, where the elements a_{ij} are given by

$$a_{11} = s; \; a_{21} = a_{31} = 0; \qquad (28)$$

$$a_{41} = k^2 + m^2 - 1 - 2\pi\mu^2\left[1 - \frac{\varepsilon-1}{2\varepsilon}\right.$$
$$\left. - \frac{(2+(\varepsilon-1)m + (\varepsilon-1)kF_m(k))\left(1 + \frac{\varepsilon-1}{2\varepsilon}[m - kG_m(k)]\right)}{(\varepsilon-1)m + \varepsilon k(F_m(k) - G_m(k))}\right];$$

$$F_m(x) \equiv \frac{I_{m+1}(x)}{I_m(x)}; \quad G_m(x) \equiv \frac{K_{m+1}(x)}{K_m(x)};$$

$$a_{12} = -(kI_{m+1}(k) + mI_m(k));$$

$$a_{22} = 2im(kI_{m+1}(k) + (m-1)I_m(k));$$

$$a_{32} = 2ik(kI_{m+1}(k) + mI_m(k));$$

$$a_{42} = [s + 2\nu(k^2 + m(m-1))]I_m(k) - 2\nu kI_{m+1}(k);$$

$$a_{13} = -imI_m(l);$$

$$a_{23} = 2lI_{m+1}(l) - (l^2 + 2m(m-1))I_m(l);$$

$$a_{33} = -mkI_m(l);$$

$$a_{43} = 2\nu im(lI_{m+1}(l) + (m-1)I_m(l));$$

$$a_{14} = -ik(lI_{m+1}(l) + mI_m(l));$$

$$a_{24} = -2mk(lI_{m+1}(l) + (m-1)I_m(l));$$

$$a_{34} = -(l^2 + k^2)(lI_{m+1}(l) + mI_m(l));$$

$$a_{44} = 2\nu ik([l^2 + m(m-1)]I_m(l) - lI_{m+1}(l)).$$

5. Evaluating explicitly the fourth-order determinant formed by the elements in (28), we obtain a dispersion equation that relates the frequencies s of jet-surface vibrations to the azimuthal number m and the wave number k; that is,

$$s^2\{m[l^2(l^2+k^2) + 2m(m-1)l^2] + F_m(l)l[(l^2+k^2)$$
$$\times(l^2 - 4m) + 2l^2m^2] - 2l^2(l^2+k^2)F_m^2(l)\}$$
$$+ 2s\nu\{-ml^2(l^2-k^2)(k^2 - m(m-1))$$
$$+ F_m(l)l[l^2k^2(l^2+k^2) + l^2m(m-1)(l^2 - 2m(m+1))$$
$$+ l^2k^2m(3m+1) - 4k^2m(k^2 - m(m^2-1))]$$
$$- 2l^2[k^2(l^2+k^2) + m(m^2-1)(l^2-k^2)]F_m^2(l)$$
$$- F_m(k)k(2l^2(l^2k^2 + m^2(m^2-1)) + l^2m(l^2 + k^2(4m-5))$$
$$+ F_m(l)l[l^4 - 5l^2k^2 + 4m(m^2-1)(l^2-k^2)]$$
$$+ 2l^2(m^2-1)(l^2-k^2)F_m^2(l))\}$$

$$= f(m, \mu, k)\{m(l^2 m(l^2 - k^2) + F_m(l) l[l^2(l^2 - k^2) - 2m(l^2 - 2k^2)] - 2l^2(l^2 - k^2)F_m^2(l)) + F_m(k)k(m[l^2(l^2 - k^2) - 2ml^2] + l(l^2 - k^2)(l^2 - 4m) \times F_m(l) - 2l^2(l^2 - k^2)F_m^2(l))\};$$

$$f(m, \mu, k) \equiv 1 - k^2 - m^2 + 2W\left\{1 - \frac{1}{2\varepsilon}\left[\varepsilon - 1 + \frac{(2 + (\varepsilon - 1)m + (\varepsilon - 1)kF_m(k))(2\varepsilon + (\varepsilon - 1)m - (\varepsilon - 1)kG_m(k))}{(\varepsilon - 1)m + \varepsilon kF_m(k) + kG_m(k)}\right]\right\}; \quad (29)$$

$$W = \pi\mu^2.$$

An analysis of Eq. (29) is very difficult in the general case and calls for the application of numerical methods.

In the limiting case of $\varepsilon \longrightarrow \infty$ and $\mu = 2\kappa$, where κ is the surface charge density of a jet formed by an electrically conducting liquid, we arrive at the dispersion equation obtained in [12] for nonaxisymmetric vibrations of a charged jet of an electrically conducting viscous liquid (in this transition, it is assumed that the charge per unit jet length remains unchanged).

The dispersion equation for axisymmetric vibrations of a charged jet—various simplified versions of this equation were discussed in [3, 5, 7]—can easily be derived from (29) by setting $m = 0$. The result is

$$s^2 + 2\nu k^2 s\left[1 - F_0(k)\left(\frac{2kl}{(l^2 + k^2)F_0(l)} + \frac{l^2 - k^2}{k(l^2 + k^2)}\right)\right] = k\frac{l^2 - k^2}{l^2 + k^2}F_0(k) f(0, \mu, k).$$

At $\mu = 0$, this equation coincides with the dispersion equation for a neutral jet of the viscous liquid [21]; that is,

$$s^2 + 2\nu k^2 s\left[1 - F_0(k)\left(\frac{2kl}{(l^2 + k^2)F_0(l)} + \frac{l^2 - k^2}{k(l^2 + k^2)}\right)\right] = k(1 - k^2)\frac{l^2 - k^2}{l^2 + k^2}F_0(k).$$

For a jet formed by a liquid of low viscosity, in which case the condition $l \gg k$ is satisfied, Eq. (29) takes a simpler form,

$$s^2 + 2s\nu(k^2 + m(m - 1) - kF_m(k)) = f(m, \mu, k)(m + kF_m(k)). \quad (30)$$

In the ideal-liquid limit ($\nu \longrightarrow 0$), Eq. (30) reduces to the form

$$s^2 = f(m, \mu, k)(m + kF_m(k)). \quad (31)$$

6. From Eq. (31), it can be seen that, for $f > 0$, the frequencies s are real-valued and that this equation has two roots. One of them is negative and determines the decrement of decay for the corresponding motion of the liquid; therefore, it is of no interest for the present investigation. The other root is positive and determines the increment of growth of the instability of a cylindrical wave, $s = \sqrt{f}\,(m + kF_m(k))^{1/2}$. Equating to zero the first derivative of the increment with respect to the wave number, one can find the wave number of the capillary wave corresponding to the maximum value of the increment; substituting thereupon this wave number into (31), one obtains the relevant increment itself. The results of such calculations based on Eq. (30) that were performed with the aid of the Matematica package for analytic calculations are given in Figs. 1 and 4 in the form of the dependences $s = s(W)$ and $k = k(W)$ for the first five values of the azimuthal number m for various values of the dielectric permittivity ε.

From a comparison of the data presented in Fig. 1 for $\varepsilon \longrightarrow \infty$, Fig. 2 for $\varepsilon = 20$, and Figs. 3 and 4 showing the results of our calculations for liquid hydrogen ($\varepsilon = 1.241$), one can easily see that, as the dielectric permittivity of liquids decreases, the absolute values of the instability increments and the wave numbers of the most unstable waves also become smaller. Moreover, the region of W values where instability is realized is shifted toward greater values of W, and this is seen most clearly from a comparison of Figs. 2–4.

The most interesting result obtained previously in [12] in studying the instability of superficially charged nonaxisymmetric jets of electrically conducting liquids is that, at rather large values of the parameter W, the increments of the instability of nonaxisymmetric modes become commensurate with the increments for axisymmetric modes, while the wave numbers of the most unstable waves fall below their counterparts for the axisymmetric case. This means that, at the final stage of the realization of instability, which is manifested in the fragmentation of a jet into drops, the drop will disintegrate into drops of different dimensions; therefore, preferential conditions must be created for the axisymmetric mode—for example, by artificially

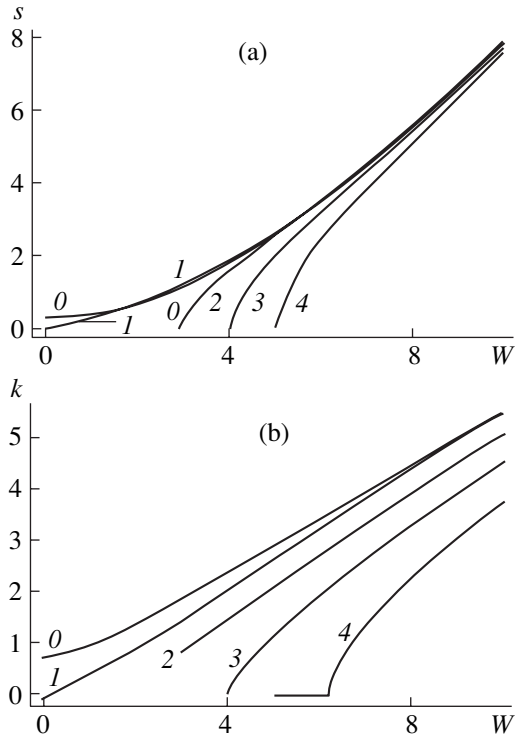


Fig. 1. (a) Dimensionless increment and (b) dimensionless wave number of the most unstable wave versus the dimensionless parameter W characterizing the surface charge according to calculations at $\nu = 0.1$ for $\varepsilon \longrightarrow \infty$. The figures on the curves correspond to the values of the azimuthal number m.

increasing its amplitude, as is done in the induced capillary decay of jets [2]—in order to solve the practical problem of obtaining fluxes of monodispersed drops.

In the above situation of spatially charged dielectric jets, the increments of the instability of nonaxisymmetric modes prove to be even greater than those for axisymmetric modes. This is seen most clearly from Figs. 3a and 4a, which display the results obtained by calculating, for a liquid-hydrogen jet, the instability increments for the axisymmetric mode (the number on a curve coincides with the number of the mode) and the next four nonaxisymmetric modes in the ascending order of the azimuthal number m. It is interesting to note that, for low values of the dielectric permittivity, the region of W values at which the wave numbers of the most unstable modes take reasonable values for $m \geq 2$ modes from the point of view of the disintegration of jets into drops in practice is shifted toward greater values of W. This is seen from a comparison of the regions of W values corresponding to the realization of jet-surface instability in Figs. 3a and 3b, as well as in Figs. 4a and 4b. In Figs. 3b and 4b, the curves have segments parallel to the abscissa (they are nearly coincident with it) that correspond to unstable waves characterized by very low wave numbers. Physically, this implies the instability of very long waves. In the sense of jet fragmentation into parts, the realization of such an instability would lead to the disintegration of a jet into large pieces rather than into small drops. Such an instability was experimentally observed in [13, 14], and

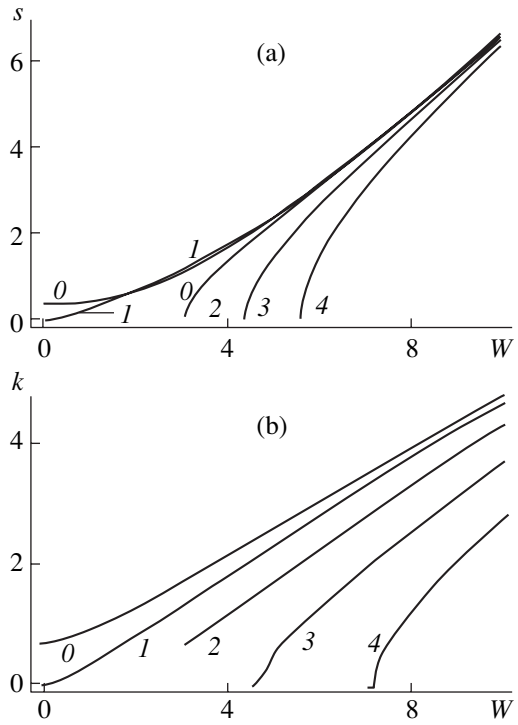


Fig. 2. As in Fig. 1, but for $\varepsilon = 20$.

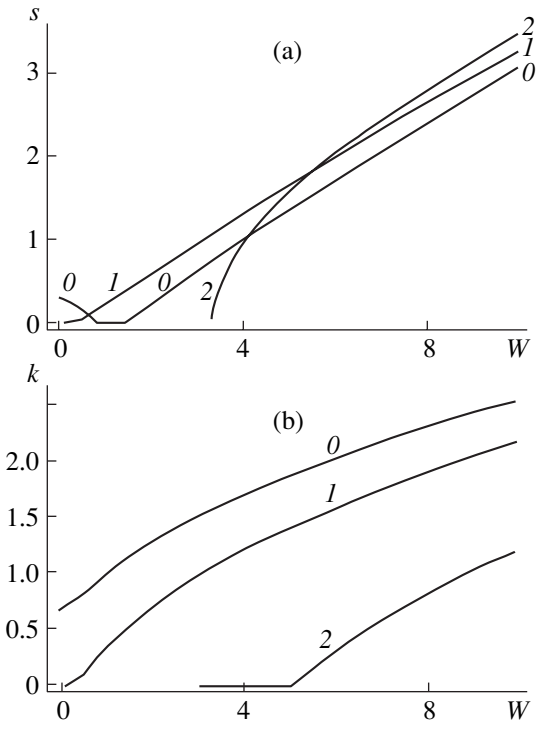


Fig. 3. As in Fig. 1, but at $m = 0$, 1, and 2 for liquid hydrogen at $\varepsilon = 1.241$.

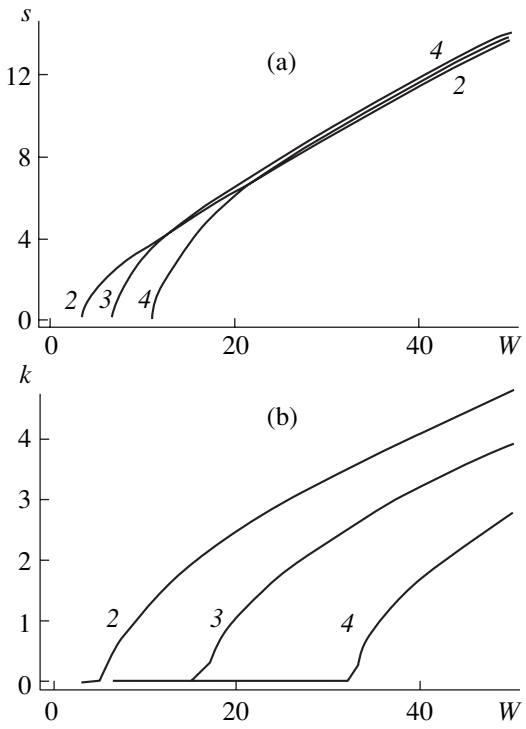


Fig. 4. As in Fig. 3, but at $m = 2$, 3, and 4 for a wider range of W values.

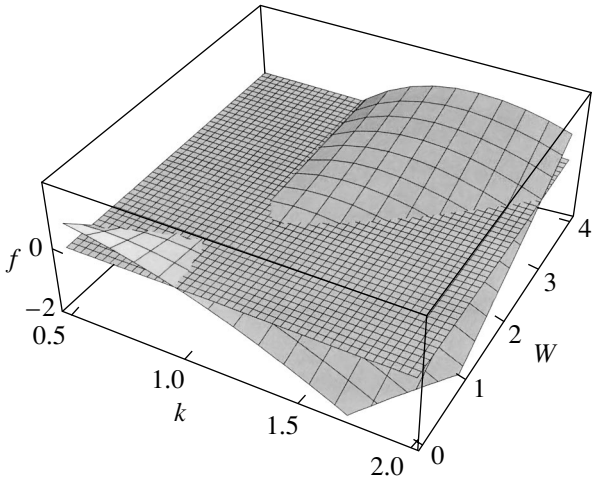


Fig. 5. Factor $f(k, m, W)$ as a function of the parameter W and the dimensionless wavenumber k (coarse grid). The fine grid shows the plane $f = 0$

this entails a significant complication of the spectrum of modes of the electrostatic dispersion of a liquid [14, 15, 22, 23].

From Fig. 3a, it can be seen that, for a jet of liquid hydrogen characterized by a rather low dielectric permittivity ($\varepsilon = 1.241$), the curve representing the dependence $s = s(W)$ for $m = 0$ first descends to zero with increasing W; after that, there are no unstable solutions to Eq. (30) in some range of W values, but they then again appear, with the result that the general form of the dependence $s = s(W)$ becomes identical to that which is displayed by the other curves. This type of behavior of the function $s = s(W)$, with a minimum in the region of low values of W, was previously observed in [12] at high values of viscosity (at $\nu = 3$) and is due to the nonmonotonic character of the function $f(k, m, W)$ at small k and W, as can be seen from Fig. 5, which shows the calculated dependence $f = f(k, m, W)$ for a dielectric jet with $\varepsilon = 1.241$ at $m = 0$. The same figure also displays the $f \equiv 0$ plane represented by a denser coordinate mesh. If one recalls that positive values of the function $f(k, m, W)$, which rise above the $f \equiv 0$ plane, correspond to unstable states of charged jets, it can be seen that, at small values of k and W, there is a locus in which unstable solutions are absent. It is this circumstance that leads to the dependence of the increment of instability of the axisymmetric ($m = 0$) mode on the parameter W in Fig. 3a.

The effect of viscosity on the regularities of realization of jet instability is qualitatively similar to that in the case of a superficially charged conducting jet, which was investigated previously in [12]: the instability increments and the wave numbers of the most unstable waves decrease with increasing viscosity. In view of this, we will not pay special attention to this issue.

CONCLUSIONS

The increments of instability of nonaxisymmetric modes of spatially charged jets become smaller with decreasing dielectric permittivity of liquids, this effect being more pronounced for smaller values of the azimuthal number m (which characterizes the degree of deviation from axial symmetry). As a result, the increments of instability of nonaxisymmetric modes in dielectric liquids of small dielectric permittivities may become substantially greater than the increment of instability of the axisymmetric mode, all other conditions being the same, and this will affect the regularities of jet fragmentation into drops.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 03-01-00760).

APPENDIX

The electrostatic-field pressure at the surface of a charged dielectric jet, P_e, is given by

$$P_e = -\mu\Phi^{in} + \frac{\varepsilon - 1}{8\pi\varepsilon}\left(\frac{\partial\Phi^{ex}}{\partial n}\right)^2 + \frac{\varepsilon - 1}{8\pi}\left(\frac{\partial\Phi^{ex}}{\partial\tau}\right)^2, \quad \text{(A1)}$$

where the potentials of the electric field within and beyond the liquid—Φ^{in} and Φ^{ex}, respectively—are

obtained by solving the boundary-value problem

$$\Delta\Phi^{in} = -4\pi\frac{\mu}{\varepsilon};\quad \Delta\Phi^{ex} = 0;$$

$$r = 1+\xi:\ \Phi^{in} = \Phi^{ex};\quad \varepsilon\frac{\partial\Phi^{in}}{\partial n} = \frac{\partial\Phi^{ex}}{\partial n}; \tag{A2}$$

$$r \longrightarrow 0:\ \Phi^{in} \longrightarrow 0.$$

We represent the potentials Φ^{in} and Φ^{ex} in the form of the expansions

$$\Phi^{in} = \Phi_0^{in} + \varphi^{in},\quad \Phi^{ex} = \Phi_0^{ex} + \varphi^{ex}, \tag{A3}$$

where Φ_0^{in} and Φ_0^{ex} are the respective electric-field potentials at the unperturbed jet surface, while φ^{in} and φ^{ex} are the first-order corrections in the perturbation $\xi(z, \phi, t)$ of the surface to these potentials. Separating the above problem according to orders of smallness, we obtain the following:
in the zeroth order,

$$\Delta\Phi_0^{in} = -4\pi\frac{\mu}{\varepsilon};\quad \Delta\Phi_0^{ex} = 0;$$

$$r \longrightarrow 0:\ \Phi_0^{in} = 0;\quad r \longrightarrow \infty:\ \varphi^{ex} \longrightarrow 0; \tag{A4}$$

$$r = 1:\ \Phi_0^{in} = \Phi_0^{ex};\quad \varepsilon\frac{\partial\Phi_0^{in}}{\partial r} = \frac{\partial\Phi_0^{ex}}{\partial r},$$

while, in the first order,

$$\Delta\varphi^{in} = 0;\quad \Delta\varphi^{ex} = 0;$$

$$r \longrightarrow 0:\ \varphi^{in} \longrightarrow 0;\quad r \longrightarrow \infty:\ \varphi^{ex} \longrightarrow 0;$$

$$r = 1:\ \varphi^{in} + \frac{\partial\Phi_0^{in}}{\partial r}\xi = \varphi^{ex} + \frac{\partial\Phi_0^{ex}}{\partial r}\xi; \tag{A5}$$

$$\varepsilon\left(\frac{\partial\varphi^{in}}{\partial r} + \frac{\partial^2\Phi_0^{in}}{\partial r^2}\xi\right) = \frac{\partial\varphi^{ex}}{\partial r} + \frac{\partial^2\Phi_0^{ex}}{\partial r^2}\xi.$$

In the first order of smallness, the potentials Φ_0^{in} and Φ_0^{ex} of the electric field within and beyond the jet are given by

$$\Phi_0^{in} = -\frac{\pi\mu r^2}{\varepsilon};\quad \Phi_0^{ex} = -\frac{\pi\mu}{\varepsilon} - 2\pi\mu\ln r. \tag{A6}$$

The solution to the first-order problem has the form

$$\varphi^{ex} = \int_0^\infty \sum_{m=0}^{\infty} C_4 K_m(kr)\exp(im\phi)\exp(ikz)\exp(st)dk;$$

$$\varphi^{in} = \int_0^\infty \sum_{m=0}^{\infty} C_5 I_m(kr)\exp(im\phi)\exp(ikz)\exp(st)dk, \tag{A7}$$

where the coefficients C_4 and C_5 are expressed in terms of the coefficient D appearing in the integral representation (24) of the perturbation function ξ (see Eq. (26)).

The distortion of the equilibrium cylindrical surface of the jet by the wave motion $\xi(z, \phi, t)$ entails a change in the pressure P_e. Since the perturbation ξ is small, we represent P_e, upon the substitution of expansions (A3) into (A1), in the form of a series in the small quantities φ^{in}, φ^{ex}, and ξ, retaining terms to the first order inclusive; that is,

$$P_e\big|_{r=1+\xi} = \left\{-\mu(\Phi_0^{in} + \varphi^{in}) + \frac{\varepsilon-1}{8\pi\varepsilon}\left(\frac{\partial\Phi_0^{ex}}{\partial n} + \frac{\partial\varphi^{ex}}{\partial n}\right)^2 + \frac{\varepsilon-1}{8\pi}\left(\frac{\partial\Phi_0^{ex}}{\partial\tau} + \frac{\partial\varphi^{ex}}{\partial\tau}\right)^2\right\}\Bigg|_{r=1+\xi} + O(\xi^2)$$

$$\approx \left\{-\mu\left(\Phi_0^{in} + \xi\frac{\partial\Phi_0^{in}}{\partial r} + \varphi^{in}\right) + \left(\frac{\varepsilon-1}{8\pi\varepsilon}\left(\frac{\partial\Phi_0^{ex}}{\partial r}\right)^2 + \frac{\varepsilon-1}{8\pi\varepsilon}\frac{\partial}{\partial r}\left(\frac{\partial\Phi_0^{ex}}{\partial r}\right)^2\xi + \frac{\varepsilon-1}{4\pi\varepsilon}\frac{\partial\Phi_0^{ex}}{\partial r}\frac{\partial\varphi^{ex}}{\partial r}\right)\right\}\Bigg|_{r=1} + O(\xi^2), \tag{A8}$$

where $\mathbf{n}$ and $\boldsymbol{\tau}$ are unit vectors along, respectively, the normal and the tangent to the jet surface.

In deriving the last expressions, we used the relations

$$\frac{\partial\Phi_0^{ex}}{\partial n} \approx \frac{\partial\Phi_0^{ex}}{\partial r} + O(\xi^2),\quad \frac{\partial\varphi^{ex}}{\partial n} \approx \frac{\partial\varphi^{ex}}{\partial r} + O(\xi^2),$$

$$\frac{\partial\Phi_0^{ex}}{\partial\tau} = 0.$$

From (A8), it can be seen that the expressions for the electric-field pressure at the unperturbed surface of a charged dielectric jet, P_e^0, and for the first-order correction in the small perturbation of the equilibrium jet surface to the electric-field pressure, p_e, are given by

$$P_e^0 \approx \left\{-\mu\Phi_0^{in} + \frac{\varepsilon-1}{8\pi\varepsilon}\left(\frac{\partial\Phi_0^{ex}}{\partial r}\right)^2\right\}\Bigg|_{r=1};$$

$$p_e \approx \left\{-\mu\left(\frac{\partial\Phi_0^{in}}{\partial r}\xi + \varphi^{in}\right) + \left(\frac{\varepsilon-1}{8\pi\varepsilon}\frac{\partial}{\partial r}\left(\frac{\partial\Phi_0^{ex}}{\partial r}\right)^2\xi + \frac{\varepsilon-1}{4\pi\varepsilon}\frac{\partial\Phi_0^{ex}}{\partial r}\frac{\partial\varphi^{ex}}{\partial r}\right)\right\}\Bigg|_{r=1}.$$

Taking into account the zeroth-order solution (A6) and relation (A5) between the potentials φ^{in} and φ^{ex} at

the jet surface, we obtain

$$P_e^0 \approx \pi\mu^2\left(\frac{1}{\varepsilon} + \frac{\varepsilon - 1}{2\varepsilon}\right);$$

$$p_e \approx 2\pi\mu^2\left(1 - \frac{\varepsilon - 1}{2\varepsilon}\right)\xi - \mu\left(\varphi^{ex} + \frac{\varepsilon - 1}{2\varepsilon}\frac{\partial\varphi^{ex}}{\partial r}\right).$$

Substituting for the functions ξ and φ^{ex} their integral representations (23) and (24) and taking into account (25) and (26), we obtain a final expression for the pressure of the electric field associated with the perturbation of the jet-surface shape. The result is

$$p_e = 2\pi\mu^2 D\left[1 - \frac{\varepsilon - 1}{2\varepsilon} - \frac{(2 + (\varepsilon - 1)kh(k))(2\varepsilon + (\varepsilon - 1)kg(k))}{2\varepsilon^2 k[h(k) - g(k)]}\right] \times \exp i(kz + m\phi)\exp(st),$$

where $h(k)$ and $g(k)$ are defined in (26).

REFERENCES

1. V. M. Entov and A. L. Yarin, Itogi Nauki Tekh., Ser.: Mekh. Zhidk. Gaza **17**, 112 (1984).
2. E. V. Ametistov, V. V. Blazhenkov, A. K. Gorodov, *et al.*, *Monodispersion of Materials: Principles and Application* (Énergoatomizdat, Moscow, 1991).
3. G. A. Glonti, Zh. Éksp. Teor. Fiz. **34**, 1329 (1958) [Sov. Phys. JETP **7**, 917 (1958)].
4. F. H. Nayfeh, Phys. Fluids **4**, 841 (1970).
5. D. A. Savill, J. Fluid Mech. **48**, 815 (1971).
6. M. Mutoh, S. Kaieda, and K. Kamimura, J. Appl. Phys. **50**, 3174 (1979).
7. S. Grossmann and A. Muller, Z. Phys. B: Condens. Matter. **57**, 161 (1984).
8. S. S. Nazin, A. N. Izotov, and V. B. Shikin, Dokl. Akad. Nauk SSSR **283**, 121 (1985) [Sov. Phys. Dokl. **30**, 606 (1985)].
9. F. J. Garcia and A. Castellanos, Phys. Fluids **6**, 2676 (1994).
10. J. Eggers, Phys. Fluids **7**, 941 (1995).
11. Yu. G. Chesnokov, Prikl. Mekh. Tekh. Fiz. **42** (3), 56 (2001).
12. S. O. Shiryaeva, A. I. Grigor'ev, T. V. Levchuk, and M. V. Rybakova, Zh. Tekh. Fiz. **73** (5), 5 (2003) [Tech. Phys. **48**, 527 (2003)].
13. W. A. Macky, Proc. R. Soc. London, Ser. A **133**, 565 (1931).
14. M. Cloupeau and B. Prunet-Foch, J. Electrost. **25**, 165 (1990).
15. S. O. Shiryaeva and A. I. Grigor'ev, J. Electrost. **34**, 51 (1995).
16. K. Kim and R. J. Turnbull, J. Appl. Phys. **47**, 1964 (1976).
17. B. I. Manzon, Usp. Fiz. Nauk **134**, 611 (1981) [Sov. Phys. Usp. **24**, 662 (1981)].
18. Y. E. Kim, M. Rabinowitz, G. S. Chulick, and R. A. Rice, Mod. Phys. Lett. B **5**, 427 (1991).
19. K. P. Pretzel, Part. World **1**, 153 (1990).
20. S. O. Shiryaeva, A. E. Lazaryants, A. I. Grigor'ev, *et al.*, Preprint No. 27 (Institute of Microelectronics, Russian Academy of Sciences, Yaroslavl, 1994).
21. V. G. Levich, *Physicochemical Hydrodynamics* (Fizmatgiz, Moscow, 1959).
22. *Handbook of Mathematical Functions*, Ed. by M. Abramowitz and L. A. Stegun (Dover, New York, 1971; Nauka, Moscow, 1979).
23. S. O. Shiryaeva and A. I. Grigor'ev, Zh. Tekh. Fiz. **64** (3), 5 (1994) [Tech. Phys. **39**, 229 (1994)].

Translated by A. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1389–1395. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 31–36.

GASES AND LIQUIDS

On the Possibility of Corona-Discharge Ignition in the Vicinity of a Weakly Charged Drop Executing Nonlinear Vibrations

A. I. Grigor'ev, S. O. Shiryaeva, and M. V. Volkova

Demidov Yaroslavl State University, Sovetskaya ul. 14, Yaroslavl, 150000 Russia

e-mail: grig@uniyar.ac.ru

Received March 20, 2003

Abstract—An analytic expression for the electric-field strength in the vicinity of a charged drop of an electrically conducting liquid is obtained for the case where the initial shape of the drop executing nonlinear vibrations is specified by a virtual excitation of an arbitrary single mode of capillary vibrations. It turns out that, even at small charges (such that the Rayleigh parameter for the drop is equal to one-tenth of the critical value associated with stability against the intrinsic charge), the electric-field strength at the drop surface in the case of an initial excitation of one of high modes is sufficient for the ignition of a corona discharge. © *2003 MAIK "Nauka/Interperiodica".*

1. According to currently prevalent concepts, the discharge of a linear lightning may be initiated by a corona discharge in the vicinity of a drop or a flooded (melting) hail freely falling in a stormy cloud [1–3]. However, measurements of charges carried by drops in stormy clouds in nature reveal that these charges are rather small, and the dimensionless Rayleigh parameter, which characterizes the stability of a drop against the intrinsic charge, does not exceed, for a drop of radius 1 mm, one-tenth of the critical value according to Rayleigh [4]. Moreover, the charges of drops are not sufficient for the strength of the electric field generated by the intrinsic charge to reach, at the spherical-drop surface, a value of $E_+ \approx 20$ kV/cm, at which the ignition of a corona discharge is possible [5]. At the same time, it is well known that the amplitude of vibrations of rainy and large ($R \approx 100$–1000 μm) cloudy drops may be as large as the drop radius itself [6]. In view of this, it seems advisable to calculate the strength of the electric field of the intrinsic charge at the vertices of a drop executing nonlinear vibrations, since it is clear from general physical considerations that the field strength must grow with increasing amplitude of vibrations.

We note that vibrations of large ($R \approx 100$–1000 μm) drops freely falling in a stormy cloud are intensified owing to their collisions with smaller drops of radius $R \sim 10$ μm, which form the bulk of the cloudy-drop concentration [4]. Thus, the vibrational energy of a drop is accumulated in high modes of its vibrations. As was shown in [7, 8], nonlinear resonance interaction in a drop of a perfect liquid (only in this model is it presently possible to perform a correct calculation of nonlinear vibrations) does not lead to vibrational-energy transfer from high modes to the lowest mode ($n = 2$), but this transfer from high modes to the third and fourth modes of vibrations proceeds quite efficiently [7, 8]. In view of the aforesaid, we will calculate the electrostatic-field strength at the surface of a drop executing nonlinear vibrations, assuming an initial excitation of one of the modes that is higher than the lowest one.

2. Suppose that we have a nonlinearly vibrating drop of a perfect, incompressible, and ideally conducting liquid of density ρ. We denote by σ the coefficient of surface tension of the liquid. We assume that the drop occurs in a vacuum and has a total charge Q; we also assume that the volume of the drop is determined by the volume of a sphere of radius R and that, by the initial instant of time $t = 0$, the equilibrium spherical shape of the drop has undergone a virtual axisymmetric perturbation of fixed amplitude ε (which is much smaller than the radius of the drop), this perturbation being specified in terms of a Legendre polynomial of order m. Below, we will use a system of dimensionless variables in which $R = p = \sigma = 1$.

In the system of spherical coordinates whose origin is taken to be coincident with the center of mass of the drop, the equation of the drop surface can be represented in the form

$$r(\Theta, t) = 1 + \xi(\Theta, t), \quad |\xi| \ll 1.$$

We assume that the motion of the liquid in the drop has a potential character and that the velocity field there is completely determined by the velocity-field potential $\psi(\mathbf{r}; t)$; that is, $\mathbf{V}(\mathbf{r}; t) = \nabla\psi(\mathbf{r}; t)$. A mathematical formulation of the problem being considered is then given by

$$\Delta\psi(\mathbf{r}; t) = 0; \quad \Delta\Phi(\mathbf{r}; t) = 0;$$

$$r \longrightarrow 0: \psi(\mathbf{r}; t) \longrightarrow 0; \quad r \longrightarrow \infty: \nabla\Phi(\mathbf{r}; t) \longrightarrow 0;$$

1063-7842/03/4811-1389$24.00

$$r = 1 + \xi(\Theta, t): \Phi = \Phi_S(t); \quad \frac{\partial \xi}{\partial t} = \frac{\partial \psi}{\partial r} - \frac{1}{r^2}\frac{\partial \xi}{\partial \Theta}\frac{\partial \psi}{\partial \Theta};$$

$$\Delta p - \frac{\partial \psi}{\partial t} - \frac{1}{2}(\nabla \psi)^2 + \frac{1}{8\pi}(\nabla \Phi)^2 = \nabla \cdot \mathbf{n};$$

$$t = 0: \xi(\Theta, t) = \xi_0 + \varepsilon P_m(\cos\Theta) \ (m \geq 2);$$

$$\frac{\partial \xi(\Theta, t)}{\partial t} = 0;$$

$$-\frac{1}{4\pi}\oint_S (\mathbf{n} \cdot \nabla \Phi) dS = Q, \quad S = \begin{cases} r = 1 + \xi(\Theta, t) \\ 0 \leq \Theta \leq \pi \\ 0 \leq \phi \leq 2\pi; \end{cases}$$

$$\int_V r^2 dr \sin\Theta d\Theta d\phi = \frac{4}{3}\pi, \quad V = \begin{cases} 0 \leq r \leq 1 + \xi(\Theta, t) \\ 0 \leq \Theta \leq \pi \\ 0 \leq \phi \leq 2\pi; \end{cases}$$

$$\int_V \mathbf{e}_r \cdot r^3 dr \sin\Theta d\Theta d\phi = 0.$$

Here, $\psi(\mathbf{r}; t)$ and $\Phi(\mathbf{r}; t)$ are the potentials of, respectively, the velocity field in the liquid and the electrostatic field; Δp is the pressure drop in the equilibrium state between the interior of the drop and the region outside it; ε is the amplitude of the initial perturbation of the drop-surface shape; and ξ_0 is a constant that is determined from the requirement of invariability of the drop volume and which, to second-order terms in the small amplitude ε inclusive, is given by

$$\xi_0 = -\varepsilon^2 \frac{1}{(2m+1)} + O(\varepsilon^3).$$

Solving the above problem by the multiscale method, as was previously done in [8–10], we obtain an analytic expression for the generatrix of the drop shape. The result is

$$\xi(\Theta, t) = 1 + \varepsilon \cos(\omega_m t) P_m(\mu)$$

$$- \varepsilon^2 \frac{1}{2}\left\{ \frac{1}{(2m+1)}[1 + \cos(2\omega_m t)] \right.$$

$$+ \sum_{j=1}^{m} [(\lambda^{(-)}_{m,m,2j} + \lambda^{(+)}_{m,m,2j})\cos(\omega_{2j} t)$$

$$\left. - (\lambda^{(-)}_{m,m,2j} + \lambda^{(+)}_{m,m,2j}\cos(2\omega_m t))] P_{2j}(\mu) \right\} + O(\varepsilon^3 t);$$

$$\mu \equiv \cos\Theta;$$

$$\omega_n^2 \equiv n(n-1)[(n+2) - W]; \quad W \equiv \frac{Q^2}{4\pi};$$

$$\lambda^{(\pm)}_{ikn} \equiv \frac{[\gamma_{ikn} \pm \omega_i \omega_k \eta_{ikn}]}{[\omega_n^2 - (\omega_i \pm \omega_k)^2]};$$

$$\gamma_{ikn} \equiv K_{ikn}\left[\omega_i^2 (n - i + 1) + 2n[k(k+1) - 1] \right.$$

$$\left. + [k(i+1) - i(2i - 2n + 7) + 3] n \frac{W}{2} \right]$$

$$+ \alpha_{ikn}\left[\frac{1}{i}\omega_i^2 + n\frac{W}{2} \right];$$

$$\eta_{ikn} \equiv K_{ikn}\left(\frac{n}{2} - i + 1 \right) + \alpha_{ikn}\frac{1}{i}\left(1 + \frac{n}{2k} \right);$$

$$K_{ikn} \equiv [C^{n0}_{n0k0}]^2;$$

$$\alpha_{ikn} \equiv -\sqrt{i(i+1)k(k+1)}\, C^{n0}_{i0k0} C^{n0}_{i(-1)k1}.$$

Here, $P_m(\mu)$ are Legendre polynomials of order m, and C^{n0}_{i0k0} and $C^{n0}_{i(-1)k1}$ are Clebsch–Gordan coefficients [11].

3. Since the problem at hand consists in calculating the strength of the electric field generated by the intrinsic charge of a drop executing nonlinear vibrations, we will write explicitly the electrostatic problems of the zeroth, first, and second orders for determining the electrostatic potential of the drop. The original electrostatic problem has the form

$$\Delta \Phi(\mathbf{r}; t) = 0; \quad \mathbf{E} = -\nabla \Phi;$$

$$r \longrightarrow \infty: \operatorname{grad}\Phi \longrightarrow 0; \quad r = 1 + \xi: \Phi(\mathbf{r}, t) = \Phi_S(t);$$

$$-\frac{1}{4\pi}\oint \nabla \Phi d\mathbf{S} = Q,$$

$$S = \{ r = 1 + \xi(\Theta, t); \quad 0 \leq \Theta \leq \pi; \quad 0 \leq \phi \leq 2\pi \}.$$

We seek a solution in the form of the expansion

$$\Phi(\mathbf{r}, t) = \Phi^{(0)}(r, \Theta, T_0, T_1) + \varepsilon \Phi^{(1)}(r, \Theta, T_0, T_1)$$
$$+ \varepsilon^2 \Phi^{(2)}(r, \Theta, T_0, T_1) + O(\varepsilon^3).$$

Substituting this expansion into the above electrostatic problem, we arrive at relevant problems of various orders in the perturbation amplitude.

In the zeroth order in ε, we have

$$\Delta \Phi^{(0)} = 0;$$

$$\nabla \Phi^0 \longrightarrow 0;$$

$$\Phi^{(0)}\big|_{r=1} = \Phi_S^{(0)};$$

$$-\frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi} \left.\frac{\partial \Phi^{(0)}}{\partial r}\right|_{r=1} \sin\Theta d\Theta d\phi = Q.$$

In the first order in ε, the problem is given by

$$\Delta\Phi^{(1)} = 0;$$

$$r \longrightarrow \infty: \nabla\Phi^{(1)} \longrightarrow 0;$$

$$\Phi^{(1)}\Big|_{r=1} = \Phi_S^{(1)}(t) - \frac{\partial\Phi^{(0)}}{\partial r}\Big|_{r=1}\xi^{(1)};$$

$$\int_0^{\pi}\int_0^{2\pi}\left(\frac{\partial\Phi^{(1)}}{\partial r}\Big|_{r=1} + \frac{\partial^2\Phi^{(0)}}{\partial r^2}\Big|_{r=1}\xi^{(1)} + 2\frac{\partial\Phi^{(0)}}{\partial r}\Big|_{r=1}\xi^{(1)}\right)\sin\Theta d\Theta d\phi = 0.$$

In the second order in ε, the problem assumes the form

$$\Delta\Phi^{(2)} = 0;$$

$$r \longrightarrow \infty: \nabla\Phi^{(2)} \longrightarrow 0;$$

$$\Phi^{(2)}\Big|_{r=1} = \Big|\Phi_S^2 - \frac{\partial\Phi^{(0)}}{\partial r}\Big|_{r=1}\xi^{(2)} - \frac{1}{2}\frac{\partial^2\Phi^{(0)}}{\partial r^2}\Big|_{r=1}(\xi^{(1)})^2 - \frac{\partial\Phi^{(1)}}{\partial r}\Big|_{r=1}\xi^{(1)};$$

$$\int_0^{\pi}\int_0^{2\pi}\left(\left(\frac{\partial^2\Phi^{(0)}}{\partial r^2}\Big|_{r=1} + 2\frac{\partial\Phi^{(0)}}{\partial r}\Big|_{r=1}\right)\xi^{(2)} + \left(\frac{1}{2}\frac{\partial^3\Phi^{(0)}}{\partial r^3}\Big|_{r=1} + 2\frac{\partial^2\Phi^{(0)}}{\partial r^2}\Big|_{r=1} + \frac{\partial\Phi^{(0)}}{\partial r}\Big|_{r=1}\right)(\xi^{(1)})^2 + \left(\frac{\partial^2\Phi^{(1)}}{\partial r^2}\Big|_{r=1} + 2\frac{\partial\Phi^{(1)}}{\partial r}\Big|_{r=1}\right)\xi^{(1)} + \frac{\partial\Phi^{(2)}}{\partial r}\Big|_{r=1} - \frac{\partial\Phi^{(1)}}{\partial r}\Big|_{r=1}\frac{\partial\xi^{(1)}}{\partial\Theta}\right)\sin\Theta d\Theta d\phi = 0.$$

Solving the above electrostatic problems by standard methods, we obtain the following expression for the electrostatic potential in the vicinity of a charged drop executing nonlinear vibrations:

$$\Phi(\mathbf{r}; t) = \frac{Q}{r} + \varepsilon Q\left[\sum_{n=2}^{\infty} M_n^{(1)}(t) + \varepsilon\sum_{n=1}^{\infty} F_n^{(2)}(t)\right]r^{-(n+1)}P_n(\mu);$$

$$F_n^{(2)}(t) \equiv M_n^{(2)}(t) + \sum_{i,k=2}^{\infty} iK_{ikn}M_i^{(1)}(t)M_k^{(1)}(t);$$

$$M_n^{(1)}(t) = \delta_{nm}\cos(\omega_n t);$$

$$M_n^{(2)}(t) = -N_n(0)\cos(\omega_n t) + N_n(t);$$

$$N_n(t) = \frac{1}{2}(\lambda_{mmn}^{(-)} + \lambda_{mmn}^{(+)}\cos(2\omega_m t)).$$

The electric-field strength $\mathbf{E}(\mathbf{r}; t) \equiv -\nabla\Phi(\mathbf{r}; t)$ in the vicinity of a charged drop executing nonlinear vibrations has the form

$$\mathbf{E}(\mathbf{r}; t) = \left\{\frac{Q}{r^2} + \varepsilon Q\left[\sum_{n=2}^{\infty} M_n^{(1)}(t) + \varepsilon\sum_{n=1}^{\infty} F_n^{(2)}(t)\right]\times(n+1)r^{-(n+2)}P_n(\mu)\right\}\mathbf{n}_r + \left\{\varepsilon Q\left[\sum_{n=2}^{\infty} M_n^{(1)}(t) + \varepsilon\sum_{n=1}^{\infty} F_n^{(2)}(t)\right]r^{-(n+2)}(1-\mu^2)^{1/2}\frac{\partial P_n(\mu)}{\partial\mu}\right\}\mathbf{n}_\Theta + O(\varepsilon^3), \quad (1)$$

where $\mathbf{n}_r$ and $\mathbf{n}_\Theta$ are unit vectors of the system of spherical coordinates.

Immediately at the drop surface, the strength component E_τ tangential to the surface of an ideally conducting liquid vanishes, while the normal component E_n has the form

$$r = 1 + \xi: E_n = Q + Q\varepsilon\left\{\sum_{n=2}^{\infty}(n-1)M_n^{(1)}(t) + \varepsilon\sum_{n=0}^{\infty}[[(n-1) - (n+1)\delta_{n0}]M_n^{(2)}(t) + \sum_{i,k=2}^{\infty}([(1-\delta_{n0})(n+1)i + 3 - (i+1)(i+2)]K_{ikn} + 0.5\alpha_{ikn}M_i^{(1)}(t)M_k^{(1)}(t)]\right\}P_n(\mu) + O(\varepsilon^3). \quad (2)$$

For various instants of time, the polar-angle dependence of the electric-field strength at the surface of a nonlinearly vibrating drop for which the Rayleigh parameter is an order of magnitude less than the critical value of $W = 0.4$ is shown in Fig. 1 according to our calculations for the cases where (a) the tenth or (b) the sixth mode is excited at the initial instant of time.

The shapes of the generatrix of a nonlinearly vibrating drop at various instants of time are displayed in Fig. 2 for the same cases of initial excitation as in Fig. 1. Figure 3 presents the electric-field strength at the drop surface as a function of time for various values of the polar angle. It is interesting to note that the field strength on the drop symmetry axis attains a maximum value after a lapse of some time rather than at the initial

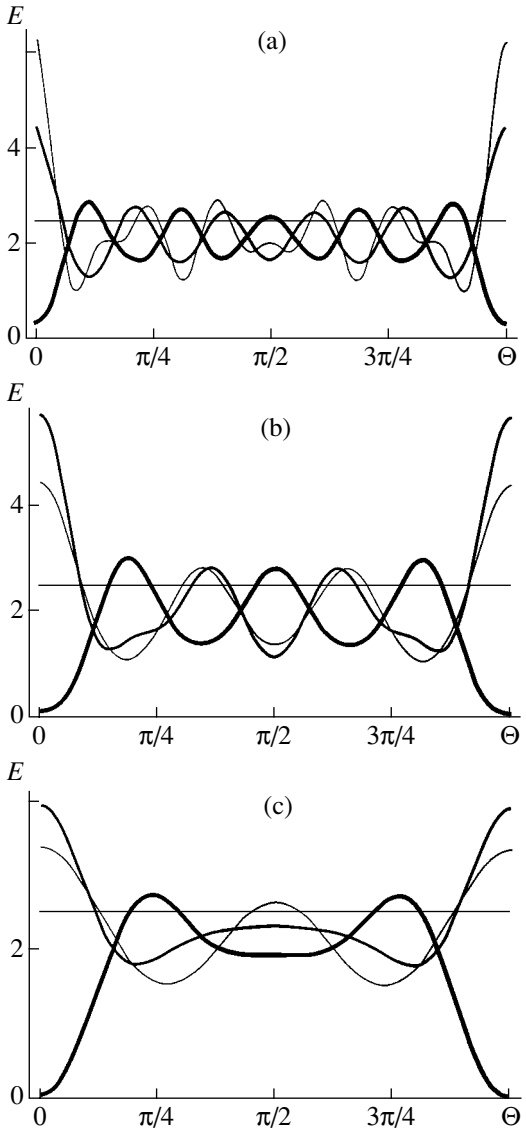


Fig. 1. Polar-angle dependences of the electric-field strength at the surface of a nonlinearly vibrating drop at $W = 0.4$ for the initial excitation of (a) the tenth mode with amplitude $\varepsilon = 0.1$ at (thin curve) $t = 0$ (initial instant), (moderately thick curve) $t = 7.0045$ (in which case the field on the symmetry axis is maximal), and (thick curve) $t = 2.4171$ (in which case the field on the symmetry axis is minimal); (b) the sixth mode with amplitude $\varepsilon = 0.1$ at the characteristic instants of time (thin curve) $t = 0$, (moderately thick curve) $t = 0.891$, and (thick curve) $t = 1.899$; and the fourth mode with amplitude $\varepsilon = 0.2$ at the characteristic instants of time (thin curve) $t = 0$, (moderately thick curve) $t = 1.26$, and (thick curve) $t = 0.1971$.

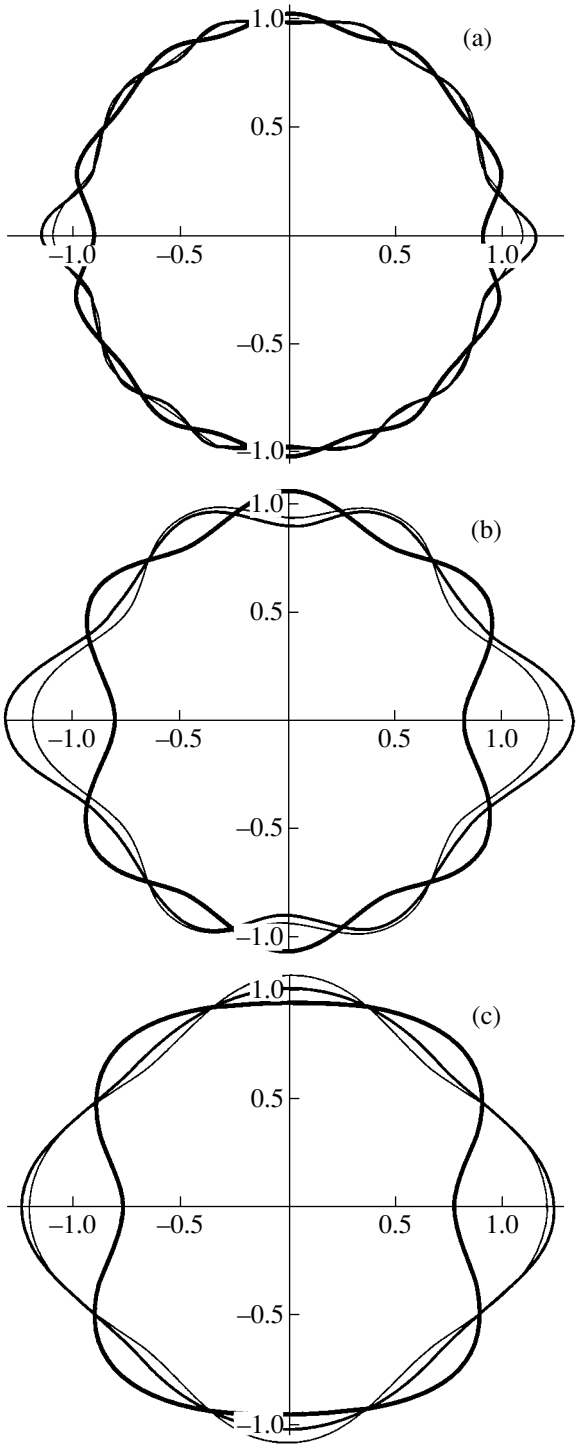


Fig. 2. Shapes of the generatrix of a nonlinearly vibrating drop at the same instants of time as in Fig. 1 for the initial excitation of the (a) tenth, (b) sixth, and (c) fourth modes.

instant; specifically, this occurs when the energy of the initial deformation is redistributed among nonlinearly excited modes in such a way that their superposition at the vertex lying on the symmetry axis ensures the minimum curvature of the vertex. It can also be seen that the characteristic time of the redistribution of energy between nonlinearly interacting modes grows with increasing number of the initial-deformation mode.

In terms of the dimensionless variables used, the critical electrostatic-field strength for the ignition of a corona discharge in the vicinity of a large water drop of radius $R = 1$ mm is about 2.5. In Figs. 1 and 3, the corresponding field-strength value is indicated by a

straight line parallel to the abscissa. It can be seen that, over a considerable part of time of nonlinear vibrations, the field strength at the vertices of the drop is sufficiently high for the ignition of a corona discharge. It is interesting to note that, in contrast to traditional interpretations [2, 11, 12] within which the possibility of corona-discharge ignition in the vicinity of a drop was associated with the growth of the amplitude of the lowest mode ($n = 2$), a local increase in the electrostatic-field strength is in fact due to the excitation of higher modes.

4. In our consideration, we disregarded the possibility of resonance energy transfer between modes, but it is well known from [8, 12, 13] that, at $W < 4$, for example, there is a degenerate three-mode resonance interaction between the fourth and sixth modes of capillary vibrations. In general, the number of resonance situations is rather large: for the first hundred vibrational modes at $W < 4$, it is as great as a few thousand [8, 14, 15]. In connection with the aforesaid, we consider the situation where the fourth mode is excited at the initial instant of time. We recall that the resonance interaction of the fourth and sixth modes is asymmetric: in the case of an initial excitation of the sixth mode, there is no resonance transfer of energy to the fourth mode, while, in the case of an initial excitation of the fourth mode, there occurs a resonance transfer of its energy to the sixth mode [13]. As a result, the evolution in the case where the initial deformation is due to a virtual excitation of the fourth mode leads to the following: the amplitudes of both the fourth and the sixth mode are of the first order of smallness, although the internal resonance interaction of the modes itself is an effect that is realized only in the second order [8].

Thus, we consider a degenerate three-mode resonance, in which case two modes are involved in a resonance interaction—that is, a relationship of the form $\omega_6 = 2\omega_4$ is valid.

Performing an analysis similar to that in [8], we obtain, for the amplitudes of the fourth and sixth modes as functions of time, expressions of the first order of smallness that have the form

$$M_6^{(1)}(t) = 2a_6^{(1)}(\varepsilon t)\cos[2\omega_4 t - \beta_6^{(1)}(\varepsilon t)]; \quad (3)$$

$$M_4^{(1)}(t) = 2a_4^{(1)}(\varepsilon t)\cos[\omega_4 t + b_4^{(1)}(\varepsilon t)], \quad (4)$$

where the functions $a_6^{(1)}(\varepsilon t)$, $\beta_6^{(1)}(\varepsilon t)$, $a_4^{(1)}(\varepsilon t)$, and $b_4^{(1)}(\varepsilon t)$ satisfy the set of differential equations

$$4\omega_6\frac{da_6^{(1)}(T_1)}{dT_1} = \Lambda_{446}^{(+)}[a_4^{(1)}(T_1)]^2\sin[\varphi_{64}^{(1)}(T_1)];$$

$$4\omega_6 a_6^{(1)}(T_1)\frac{d\beta_6^{(1)}(T_1)}{dT_1} = 4\omega_6^2 a_6^{(1)}(T_1)\delta$$

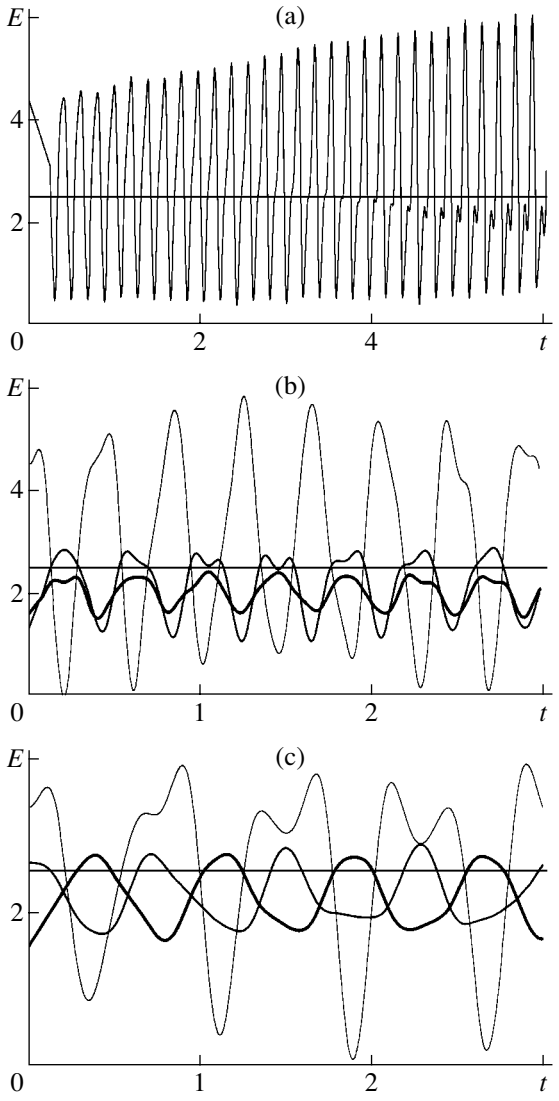


Fig. 3. Time dependences of the field strength at the surface of a drop at the same parameter values as in Fig. 1 for the initial excitation of (a) the tenth mode (field-strength values at the drop symmetry axis, $\Theta = 0$; the amplitude of vibrations reaches a maximum beyond the graph at $t = 7.0045$); (b) the sixth mode at (thin curve) $\Theta = 0$, (moderately thick curve) $\Theta = \pi/2$, and (thick curve) $\Theta = \pi/4$; (c) the fourth mode at (thin curve) $\Theta = 0$, (moderately thick curve) $\Theta = \pi/2$, and (thick curve) $\Theta = \pi/4$.

$$+ \Lambda_{446}^{(+)}[a_4^{(1)}(T_1)]^2\cos[\varphi_{64}^{(1)}(T_1)];$$

$$2\omega_4\frac{da_4^{(1)}(T_1)}{dT_1} = -\Lambda_{644}^{(-)}a_6^{(1)}(T_1)a_4^{(1)}(T_1)\sin[\varphi_{64}^{(1)}(T_1)];$$

$$2\omega_4 a_4^{(1)}(T_1)\frac{db_4^{(1)}(T_1)}{dT_1} \quad (5)$$

$$= -\Lambda_{644}^{(-)}a_6^{(1)}(T_1)a_4^{(1)}(T_1)\cos[\varphi_{64}^{(1)}(T_1)];$$

$$\varphi_{64}^{(1)}(T_1) = \beta_6^{(1)}(T_1) + 2b_4^{(1)}(T_1);$$

$$\Lambda_{ikn}^{(\pm)} = (\gamma_{ikn} + \gamma_{kin}) \pm \omega_i\omega_k(\eta_{ikn} + \eta_{kin});$$

$$a_6^{(1)}(0) = 0; \quad \beta_6^{(1)}(0) = \pi/2; \quad a_4^{(1)}(0) = 1/2;$$

$$b_4^{(1)}(0) = 0; \quad T_1 \equiv \varepsilon t,$$

with δ being a detuning parameter that is determined from the relation

$$2\omega_4 = \omega_6(1 + \varepsilon\delta).$$

The relations for amplitudes of the second order of smallness are derived in the same way as in the absence of a resonance [8]. The expressions for the electric-field strength in the vicinity of such a drop in the presence of a resonance transfer of energy between the fourth and the sixth mode will still have the form specified by Eqs. (1) and (2), but the coefficients $M_n^{(1)}(t)$ there will take the form given by (3)–(5).

The results of the calculations based on Eqs. (2) and (3)–(5) and performed for the situation under analysis are given in Figs. 1c, 2c, and 3c. There are no qualitative changes in relation to the data in Figs. 1a, 1b, 2a, 2b, 3a, and 3b for the situation where there is no resonance transfer of energy between the modes, so that all of the above conclusions on the possibility of corona-discharge ignition in the vicinity of a drop executing nonlinear vibrations remain in force.

From a comparison of curves in Figs. 1 and 3, one can see that, irrespective of the presence or absence of an internal nonlinear resonance interaction of modes, the electric-field strength at the drop surface grows with increasing number of the mode that determines the initial deformation of the equilibrium spherical shape of the drop. Figure 2 also shows that an increase in the number of the initially excited mode leads to the growth of the curvature of outliers at the drop surface. Since the charge of the drop is set to the same value in all cases, the obvious conclusion from the above is that it is the growth of the curvature of the surface in the case of the excitation of high modes that leads to an increase in the field strength at the surface of the drop executing nonlinear vibrations.

Under actual conditions of a stormy cloud, nonlinear vibrations of freely falling large ($R \approx 100$–1000 μm) drops are excited owing to their collisions with smaller drops of radius $R \sim 10$ μm, which correspond to the maximum of the size distribution of cloudy drops [4]. In such collisions, nonlinear vibrations of modes whose numbers m lie in the range between 10 and 100 are excited in large drops. If a drop carries an electric charge such that the Rayleigh parameter W reaches one-tenth of the critical value with respect to the instability of the lowest vibrational mode, $W_{cr} = 4$, the strength of the electric field generated by this charge becomes sufficiently for corona-discharge ignition in the vicinity of the drop.

An analysis of expressions (1) and (2) for the field strength at the surface of a drop executing nonlinear vibrations reveals that, as the number of the mode responsible for the initial deformation is increased, the asymptotic expansion for E becomes uniform at ever smaller values of the small parameter ε. By way of example, we indicate that, in the case where the fourth mode is initially excited, the expansions in (1) and (2) are asymptotic for $\varepsilon \le 0.2$, while, for the initial excitations of the sixth and tenth modes, this is so only for $\varepsilon \le 0.1$. The reason for this is that the energy introduced in the vibrating system by an initial deformation of the nth mode with an amplitude ε is proportional to n^2, all other conditions being the same. At rather large values of ε, the redistribution of this energy between modes that are excited in the second order of smallness owing to a nonlinear interaction leads to a considerable growth of amplitudes for some modes of low number and to a breakdown of uniformity in the relevant expansions. A significant increase in the electrostatic-field strength at the vertices of the drop with increasing number of the mode determining the initial deformation (this is illustrated by the figures in the article above) is also due to the increase in the energy introduced in the vibrating system.

In the above arguments, we did not take into account the effect of an external electrostatic field that is always present in a stormy cloud, reaching a value of about 10 kV/cm [4], and which polarizes drops. The emergence of an additional polarization charge in a drop may lead to a significant increase in the field strength in its vicinity.

CONCLUSIONS

The electric-field strength at the vertices of a weakly charged drop executing nonlinear vibrations may exceed the critical value for corona-discharge ignition in the vicinity of this drop and becomes greater as the number of the mode responsible for the initial deformation of the spherical shape of the drop is increased.

REFERENCES

1. V. A. Dyachuk and V. M. Muchnik, Dokl. Akad. Nauk SSSR **248**, 60 (1979).
2. M. N. Beĭtuganov, Meteorol. Gidrol., No. 9, 42 (1989).
3. A. I. Grigor'ev and S. O. Shiryaeva, Phys. Scr. **54**, 660 (1996).
4. *Clouds and Cloud Atmosphere: A Handbook*, Ed. by I. P. Mazina, A. Kh. Khrgian, and I. M. Imyanitov (Gidrometeoizdat, Leningrad, 1989).
5. E. D. Lozanskiĭ and O. B. Firsov, *Theory of Spark* (Atomizdat, Moscow, 1975).
6. K. V. Beard, Rev. Geophys. **25**, 357 (1987).
7. S. O. Shiryaeva, D. F. Belonozhko, and A. I. Grigor'ev, Pis'ma Zh. Tekh. Fiz. **28** (22), 45 (2002) [Tech. Phys. Lett. **28**, 945 (2002)].

8. S. O. Shiryaeva, Zh. Tekh. Fiz. **73** (2), 19 (2003) [Tech. Phys. **48**, 152 (2003)].
9. S. O. Shiryaeva, Zh. Tekh. Fiz. **71** (2), 27 (2001) [Tech. Phys. **46**, 158 (2001)].
10. S. O. Shiryaeva, Izv. Ross. Akad. Nauk, Mekh. Zhidk. Gaza, No. 3, 173 (2001).
11. D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskiĭ, *Quantum Theory of Angular Momentum* (Nauka, Leningrad, 1975; World Sci., Singapore, 1988).
12. J. A. Tsamopoulos and R. A. Brown, J. Fluid Mech. **147**, 373 (1984).
13. S. O. Shiryaeva, Pis'ma Zh. Tekh. Fiz. **26** (22), 76 (2000) [Tech. Phys. Lett. **26**, 1016 (2000)].
14. S. O. Shiryaeva, A. I. Grigor'ev, and D. F. Belonozhko, Pis'ma Zh. Tekh. Fiz. **29** (6), 69 (2003) [Tech. Phys. Lett. **29**, 249 (2003)].
15. A. N. Zharov, S. O. Shiryaeva, and A. I. Grigor'ev, Pis'ma Zh. Tekh. Fiz. **29** (9), 75 (2003) [Tech. Phys. Lett. **29**, 388 (2003)].

Translated by A. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1396–1406. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 37–46.

GASES AND LIQUIDS

Nonlinear Periodic Waves at the Charged Surface of an Electrically Conducting Viscous Liquid

D. F. Belonozhko and A. I. Grigor'ev

Yaroslavl State University, Sovetskaya ul. 14, Yaroslavl, 150000 Russia
e-mail: grig@uniyar.ac.ru

Received March 20, 2003

Abstract—A correct solution to the problem of periodic-wave propagation along the charged surface of a deep viscous liquid in the second-order approximation in the wave amplitude is given for the first time. It is shown that the second-order correction in the amplitude to the profile of the wave being considered plays a decisive role in the realization of the instability of a liquid with respect to its intrinsic charge. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The theory of waves at the surface of a liquid is one of the oldest and well-developed fields of classical fluid dynamics. The earliest achievements in this field are compiled in the textbook of Lamb [1]. As far back as the beginning of the twentieth century, there arose the trend toward exploring nonlinear effects associated with wave motion. The first successful investigations of nonlinear waves were performed for a perfect liquid [2, 3]. Studies devoted in that period to the effect of viscosity on nonlinear waves [4, 5] were less fundamental—they contained inaccuracies and even errors. Because of the presentation in terms of coordinates, which was adopted at that time in fluid dynamics, a detailed account of the full mathematical formulation of the problem of nonlinear waves in a viscous liquid (it is rather cumbersome even in modern notation) could not be given, along with a clear explanation of the essence of key points in constructing relevant solutions, within the scope of a scientific article in a journal. In view of this, the opinion then prevalent among the scientific community was that investigations of nonlinear waves in the perfect-liquid approximation provide adequate information about special features of a nonlinear wave motion; moreover, such investigations made it possible to obtain unique results by using much more modest means, since these were actually the first steps in nonlinear physics. As a result, it became common practice to give priority to the perfect-liquid approximation in studying the properties of nonlinear waves at the surface of a deep liquid.

The avalanche of articles devoted to analytic investigations of waves at the surface of a perfect liquid (see, for example, [1–13]) considerably exceeds the number of attempts at constructing a correct analytic description of the effect of viscosity on the time evolution of a nonlinear surface wave. The majority of studies dealing with waves in a viscous liquid were performed in the approximation of low viscosity within the boundary-layer theory [14–16], which is valid only at large values of the Reynolds number. Only in recent years did there appear studies [17, 18] in which a correct expression for the profile of a nonlinear periodic wave at the surface of a deep liquid of finite viscosity was found correctly in the second-order approximation in the wave amplitude.

Following [17, 18], we propose here a rigorous asymptotic solution to the problem of determining the profile of a nonlinear periodic wave that propagates along the charged surface of a deep liquid of arbitrary viscosity. This solution, which was not known in the literature until the present time, is of considerable interest for various applications of the electric dispersion of a liquid [19].

MATHEMATICAL FORMULATION OF THE PROBLEM

We assume that, in the system of Cartesian coordinates with the z axes directed vertically upward in the gravity-force field, $\mathbf{g} \parallel -\mathbf{e}_z$, an incompressible Newtonian liquid characterized by a kinematic viscosity ν, a density ρ, and the coefficient of surface tension γ fills the half-space $z \le 0$ and borders on a vacuum. We also assume that this liquid is an ideal conductor carrying a uniformly distributed surface charge such that, in the limit $z \longrightarrow \infty$, the electric field above the liquid surface distorted by wave motion tends to a uniform field of strength $E_0\mathbf{e}_z$. We will investigate the time evolution of the initial deformation of a free liquid surface.

We denote by $u = u(x, z, t)$ and $v = v(x, z, t)$, respectively, the horizontal and the vertical component of the velocity field of wave motion in the liquid and assume, for the sake of simplicity, that they are independent of the coordinate y; we also denote by $\mathbf{e}_x$ and $\mathbf{e}_z$ the unit vectors along the x and z axes, respectively. The deviation $\xi = \xi(x, t)$ of the free liquid surface from the equi-

1063-7842/03/4811-1396$24.00

librium shape $z = 0$ in the gravity-force field, the velocity field $\mathbf{U} = u \cdot \mathbf{e}_x + v \cdot \mathbf{e}_z$, and the electric potential Φ above the liquid then satisfy the relations

$$z > \xi: \Delta\Phi = 0; \tag{1}$$

$$z < \xi: \partial_t \mathbf{U} + (\nabla \cdot (\mathbf{U})) \times \mathbf{U} = -\nabla\left(\frac{1}{\rho}p + \frac{U^2}{2} + gz\right) + \nu\Delta\mathbf{U}; \tag{2}$$

$$\nabla \cdot \mathbf{U} = 0; \tag{3}$$

$$z = \xi: \partial_t \xi + u\partial_x \xi = v; \tag{4}$$

$$p - 2\rho\nu\mathbf{n}((\mathbf{n} \cdot \nabla)\mathbf{U}) + \frac{1}{8\pi}(\nabla\Phi)^2 = \gamma\nabla \cdot \mathbf{n}; \tag{5}$$

$$\boldsymbol{\tau}((\mathbf{n} \cdot \nabla)\mathbf{U}) + \mathbf{n}((\boldsymbol{\tau} \cdot \nabla)\mathbf{U}) = 0; \tag{6}$$

$$\Phi = 0; \tag{7}$$

$$z \longrightarrow +\infty: -\nabla\Phi \longrightarrow E_0\mathbf{e}_z; \tag{8}$$

$$z \longrightarrow -\infty: \mathbf{U} \longrightarrow 0. \tag{9}$$

Here, t is time; p is the pressure within the liquid; ∂_t and ∂_x are the partial derivatives with respect to time and the coordinate x, correspondingly; and $\boldsymbol{\tau}$ and $\mathbf{n}$ are the unit vectors along, respectively, the tangent and the normal to the surface. The explicit expressions for these unit vectors are given in Appendix A, along with the expression for the divergence of the normal, $\nabla \cdot \mathbf{n}$.

In order to close up the mathematical formulation of the problem, the above relations must be supplemented with initial conditions that preset the initial deviation of the surface and the initial velocity field. In just the same way as in [18], the initial conditions will be specified in solving the problem; this will be done in such a way as to arrive at the simplest (in the sense of an analytic description) forms of free-surface vibrations and to reveal the properties of a wave that are associated with its nonlinearity.

METHOD FOR SOLVING THE PROBLEM

Suppose that the initial perturbation $\xi(x, t)$ is periodic in x and that it forms a wavy relief of wavelength $\lambda = 2\pi/k$ (k is the wave number) and amplitude a. As a small parameter, we take $\varepsilon = ka$.

In the second-order approximation in ε, the solution to the problem specified by Eqs. (1)–(9) is sought in the form of the expansions

$$\begin{gathered}\mathbf{U} = \mathbf{U}_1 + \mathbf{U}_2 + O(\varepsilon^3); \quad \mathbf{U}_1 \sim O(\varepsilon); \quad \mathbf{U}_2 \sim O(\varepsilon^2);\\ p = p_0 + p_1 + p_2 + O(\varepsilon^3);\\ p_0 \sim O(1); \quad p_1 \sim O(\varepsilon); \quad p_2 \sim O(\varepsilon^2);\\ \Phi = \Phi_0 + \Phi_1 + \Phi_2 + O(\varepsilon^3);\\ \Phi_0 \sim O(1); \quad \Phi_1 \sim O(\varepsilon); \quad \Phi_2 \sim O(\varepsilon^2);\\ \xi = \xi_1 + \xi_2 + O(\varepsilon^3); \quad \xi_1 \sim O(\varepsilon); \quad \xi_2 \sim O(\varepsilon^2).\end{gathered} \tag{10}$$

Upon substituting (10) into (1)–(3), we arrive at the zeroth-, first-, and second-order problems in the above small parameter. In Appendix B, we present relations that arise upon separating the boundary conditions (4)–(7) at the free surface into relations for quantities of different order of smallness. A detailed derivation of these relations can be found in [18].

In the zeroth-order approximation in ε, the problem reduces to determining the distribution of the hydrostatic pressure in the liquid,

$$u_0 = 0; \quad v_0 = 0; \quad p_0 = -\frac{E_0^2}{8\pi} - \rho g z; \quad \Phi_0 = -E_0 z.$$

As in [18], we will use below a special notation for linear differential operators,

$$\mathscr{L} \equiv \begin{bmatrix} \partial_t - \nu(\partial_{xx} + \partial_{zz}) & 0 & (1/\rho)\partial_x & 0 \\ 0 & \partial_t - \nu(\partial_{xx} + \partial_{zz}) & (1/\rho)\partial_z & 0 \\ \partial_x & \partial_z & 0 & 0 \\ 0 & 0 & 0 & \partial_{xx} + \partial_{zz} \end{bmatrix};$$

$$\mathscr{R} \equiv \begin{bmatrix} \partial_t \\ -\rho g + \gamma\partial_{xx} \\ 0 \\ -E_0 \end{bmatrix};$$

$$\mathscr{B} \equiv \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & -2\rho\nu\partial_z & 1 & -E_0/(4\pi)\partial_z \\ \partial_z & \partial_x & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{z=0},$$

and for matrix-columns with relevant conventions for constituent elements,

$$\hat{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \quad \hat{Y}_j \equiv \begin{bmatrix} u_j \\ v_j \\ p_j \\ \Phi_j \end{bmatrix}; \quad \begin{aligned} Y_j[1] &\equiv u_j; \\ Y_j[2] &\equiv v_j; \\ Y_j[3] &\equiv p_j; \\ Y_j[3] &\equiv p_j. \end{aligned}$$

In applying the operator $\mathscr{B}$ to objects of the Y_j type, one successively performs matrix operations, differentiations, and arithmetic operations and then sets $z = 0$. The result of applying the operator $\mathscr{B}$ to a column of four functions depending on the variables x, z, and t is a column of four functions independent of z.

The above special notation was introduced in order to ensure, without violating space limitations for an

article in a journal, the clarity of presentation of key steps in solving the problem.

In the first-order approximation, we obtain a homogeneous linear differential problem, while, in the second-order approximation, we are dealing with the same linear differential problem in the presence of a nonhomogeneity whose form is determined in terms of the solutions derived in the zeroth- and first-order approximations.

FIRST-ORDER PROBLEM

For quantities of the first order in ε, the complete mathematical formulation of the problem in the above notation has the form

$$\mathcal{L}\hat{Y}_1 = \hat{0}; \tag{11}$$

$$\mathcal{B}\hat{Y}_1 + \mathcal{R}\xi_1 = \hat{0}; \tag{12}$$

$$z \longrightarrow +\infty: \Phi_1 \equiv Y_1[4] \longrightarrow 0; \tag{13}$$

$$z \longrightarrow -\infty: u_1 \equiv Y_1[1] \longrightarrow 0; \quad v_1 \equiv Y_1[2] \longrightarrow 0; \tag{14}$$

$$\begin{gathered} t = 0: \xi_1 = a\cos(kx); \\ z \le 0: Y_1[1]_{t=0} = u_1^0; \quad Y_1[2]_{t=0} = v_1^0. \end{gathered} \tag{15}$$

For an initial deformation in the first approximation, we chose a cosine function for the sake of simplicity. The functions $u_1^0 \equiv u_1^0(x, z)$ and $v_1^0 \equiv v_1^0(x, z)$ describing the initial velocity distribution will be specified below. We note that initial conditions must be formulated only for the first two elements of $\hat{Y}_1$ rather than for all elements of this column of unknown quantities. For p_1 and Φ_1, one needs no initial conditions. For p_1, this fact can be explained as follows: The linearized Navier—Stokes equation (the first equation in the set of Eq. (11)) can be recast into the form

$$\partial_t \mathbf{U}_1 = -\nabla\left(\frac{1}{\rho}p_1\right) + \nu\Delta\mathbf{U}_1.$$

Applying the div operation to both sides of this equation and considering that the liquid is incompressible (that is, $\nabla \cdot \mathbf{U}_1 = 0$) and that the successive partial derivatives are commutative, we can easily find that p_1 satisfies Laplace's equation

$$\Delta p_1 = 0.$$

The linearized boundary condition for the normal tensions and the condition that the gradient of the first-order correction to pressure vanishes at a large depth have the form

$$z = 0: p_1 = 2\rho\nu\partial_z v_1 + \partial_{xx}\xi_1; \quad z \longrightarrow -\infty: |\nabla p_1| \longrightarrow 0.$$

It can be seen that, if the velocity field is known at the surface and within the interior of the liquid and if the expression for ξ_1 is preset, the pressure p_1 satisfies the Dirichlet problem within an unbounded region, the solution to this problem existing and being unique. No initial condition is imposed on Φ_1 for a similar reason. The aforesaid can be qualitatively explained as follows: within the model used here (that of an ideally conducting incompressible liquid), any variation in the velocity field and in the perturbation of the surface, ξ_1, instantaneously affects the form of the scalar fields p_1 and Φ_1.

For the first-order problem, the solution described by relations that differ from the analogous ones constructed in [18] only by the dispersion equation where there appears a term proportional to the square of the electrostatic-field strength, E_0^2, can easily be obtained by traditional methods [20]. This yields

$$\xi_1^* = a\cos(\theta)\exp(T); \tag{16}$$

$$\begin{aligned} u_1^* = a((&S_2\exp(kz) - 2\nu k(q_2\cos(q_2 z) + q_1\sin(q_2 z)) \\ &\times \exp(q_1 z))\cos(\theta) + (D\exp(kz) - 2\nu k(q_1\cos(q_2 z) \\ &- q_2\sin(q_2 z))\exp(q_1 z))\sin(\theta))\exp(T); \end{aligned} \tag{17}$$

$$\begin{aligned} v_1^* = a((&D\exp(kz) - 2\nu k^2\cos(q_2 z)\exp(q_1 z))\cos(\theta) \\ &- (S_2\exp(kz) - 2\nu k^2\sin(q_2 z)\exp(q_1 z))\sin(\theta))\exp(T); \end{aligned} \tag{18}$$

$$\begin{aligned} p_1^* = a\rho k^{-1}((&-S_1 D + S_2^2)\cos(\theta) \\ &+ 2S_2(S_1 + \nu k^2)\sin(\theta))\exp(kz)\exp(T); \end{aligned} \tag{19}$$

$$\Phi_1^* = aE_0\cos(\theta)\exp(T); \tag{20}$$

$$\nu^2(k^2 + q^2)^2 - 4\nu^2 k^3 q + k\left(g + \frac{k^2\gamma}{\rho} - \frac{k}{\rho}\frac{E_0^2}{4\pi}\right) = 0; \tag{21}$$

$$q_1 = \mathrm{Re}\, q \ge 0; \quad q_2 = \mathrm{Im}(q) \ge 0; \tag{22}$$

$$S = \nu(q^2 - k^2); \quad S_1 = \mathrm{Re}S; \quad S_2 = \mathrm{Im}S; \tag{23}$$

$$\theta = S_2 t - kx; \quad T = S_1 t; \quad D = S_1 + 2\nu k^2. \tag{24}$$

In these relations, the quantity q is calculated as that root of the dispersion equation (21) which satisfies conditions (22). As in [18], these are conditions for selecting the root that corresponds to wave motion at a velocity that vanishes for $z \longrightarrow -\infty$ (see condition (14)) and to which a progressive wave propagating to the right corresponds. Conditions (22) ensure the uniqueness of the procedure for calculating the complex frequency S.

With the aid of the solution specified by Eqs. (16)–(20), the substitution

$$\xi = \xi_1^* + \xi_1^{**}; \quad Y_1 = Y_1^* + Y_1^{**}$$

reduces (11)–(15) to a problem where there is no initial perturbation of the surface; that is,

$$\mathcal{L}\hat{Y}_1^{**} = \hat{0}; \quad \mathcal{B}\hat{Y}_1^{**} + \mathcal{R}\xi_1^{**} = \hat{0};$$

$$z \longrightarrow +\infty: \Phi_1^{**} \equiv Y_1[4] \longrightarrow 0;$$

$$z \longrightarrow -\infty: u_1^{**} \equiv Y_1[1] \longrightarrow 0; \quad v_1^{**} \equiv Y_1[2] \longrightarrow 0; \quad (25)$$

$$t = 0: \xi_1^{**} = 0; \quad z \le 0: Y_1^{**}[1]_{t=0} = u_1^0 - Y_1^*[1]_{t=0};$$
$$Y_1^{**}[2]_{t=0} = v_1^0 = Y_1^*[2]_{t=0}.$$

The shape of a free surface is represented as a superposition of the function ξ_1^*, which, at $t = 0$, coincides with the initial distortion of the surface, and the function ξ_1^{**}, which, at the initial instant, coincides with the equilibrium surface ($\xi_1^{**} = 0$ at $t = 0$). In order to obtain the least cumbersome solution, we assume that, in the problem described by Eq. (25), the velocity of all points of the liquid at the initial instant satisfies the conditions

$$z \le 0: u_1^0 - Y_1^*[1]_{t=0} = 0; \quad v_1^0 - Y_1^*[2]_{t=0} = 0. \quad (26)$$

It follows that the solution to the problem specified by Eq. (25) is trivial and that relations (16)–(24) provide the solution to the first-order problem with the initial condition (26).

SECOND-ORDER PROBLEM

Owing to isolating relations for quantities of different orders in (1)–(9) (see Appendix B) and subsequently substituting the solution obtained in the first-order approximation into the resulting expressions dependent on first-order quantities, the second-order problem with initial conditions that have yet to be specified can be formulated as

$$\mathcal{L}\hat{Y}_2 = a^2\mathrm{Re}\Bigg(\Bigg(\begin{bmatrix}\hat{A}_1\\0\end{bmatrix}\exp(2q_1z) + \begin{bmatrix}\hat{A}_2\\0\end{bmatrix}\exp(2kz) + \begin{bmatrix}\hat{A}_3\\0\end{bmatrix}\exp((k+q)z)\Bigg)\exp(2T) + \begin{bmatrix}\hat{A}_4\\0\end{bmatrix}\exp((k+q)z)\exp(2(T+i\theta))\Bigg); \quad (27)$$

$$\mathcal{B}\hat{Y}_2 + \mathcal{R}\xi_2 = a^2\mathrm{Re}\Bigg(\begin{bmatrix}\hat{A}_5\\\frac{1}{2}kE_0\end{bmatrix}\exp(2T) + \begin{bmatrix}\hat{A}_6\\\frac{1}{2}kE_0\end{bmatrix}\exp(2(T+i\theta))\Bigg); \quad (28)$$

$$z \longrightarrow +\infty: \Phi_2 \equiv Y_2[4] \longrightarrow 0; \quad (29)$$

$$z \longrightarrow -\infty: u_2 \equiv Y_2[1] \longrightarrow 0; \quad v_2 \equiv Y_2[2] \longrightarrow 0, \quad (30)$$

where $\hat{A}_1$, $\hat{A}_2$, $\hat{A}_3$, $\hat{A}_4$, $\hat{A}_5$, and $\hat{A}_6$ are three-element columns with complex-valued coefficients that do not depend on the coordinates and time.

The expressions for these columns are given in Appendix C.

Since the above problem is linear, we can find its complex-valued solution and then obtain, after taking its real part, the physical solution. Following the method of solution described in detail in [18], we find that the sets of nonhomogeneous equations

$$\mathcal{L}\hat{y}_a = a^2\Bigg(\begin{bmatrix}\hat{A}_1\\0\end{bmatrix}\exp(2q_1z) + \begin{bmatrix}\hat{A}_2\\0\end{bmatrix}\exp(2kz) + \begin{bmatrix}\hat{A}_3\\0\end{bmatrix}\exp((k+q)z)\Bigg)\exp(2T); \quad (31)$$

$$\mathcal{L}\hat{y}_b = a^2\begin{bmatrix}\hat{A}_4\\0\end{bmatrix}\exp((k+q)z)\exp(2(T+i\theta)), \quad (32)$$

which are formulated for unknown quantities that form the columns $\hat{y}_a$ and $\hat{y}_b$, have the particular solutions

$$\hat{y}_a = a^2\Bigg(\begin{bmatrix}\hat{C}_1\\0\end{bmatrix}\exp(2q_1z) + \begin{bmatrix}\hat{C}_2\\0\end{bmatrix}\exp(2kz) + \begin{bmatrix}\hat{C}_3\\0\end{bmatrix}\exp((k+q)z)\Bigg)\exp(2T); \quad (33)$$

$$\hat{y}_b = a^2\begin{bmatrix}\hat{C}_4\\0\end{bmatrix}\exp((k+q)z)\exp(2(T+i\theta)), \quad (34)$$

where $\hat{C}_1$, $\hat{C}_2$, $\hat{C}_3$, and $\hat{C}_4$ are three-element columns with coefficients that have yet to be determined.

A direct substitution of the solutions given by (33) and (34) into Eqs. (31) and (32) leads to a set of nonhomogeneous linear equations for the coefficients of the columns $\hat{C}_i$. For them, we obtain

$$C_1 = \hat{\Pi}_1^{-1}\hat{A}_1; \quad C_2\hat{\Pi}_2^{-1}\hat{A}_1;$$
$$C_3 = \hat{\Pi}_3^{-1}\hat{A}_3; \quad C_4 = \hat{\Pi}_4^{-1}\hat{A}_4, \quad (35)$$

where $\hat{\Pi}_1$, $\hat{\Pi}_2$, $\hat{\Pi}_3$, and $\hat{\Pi}_4$ are square matrices with coefficients that are independent of the coordinates and time and which are presented in Appendix C.

Further, we use standard methods as in [18] to seek solutions to the homogeneous problem

$$\mathcal{L}(\hat{y}_c) = 0 \quad (36)$$

in the form

$$\hat{y}_c = a^2\begin{bmatrix}0\\0\\f(t)\\\mathcal{F}(t)\end{bmatrix}\exp(2T) + a^2\left(2\begin{bmatrix}-ik\\k\\-\rho S\\0\end{bmatrix}\lambda\exp(2kz)\right.$$

$$+\begin{bmatrix}-r\\-2ik\\0\\0\end{bmatrix}H\exp(z\sqrt{2(k^2+q^2)}) \qquad (37)$$

$$\left.+\begin{bmatrix}0\\0\\0\\1\end{bmatrix}R\exp(-2kz)\right)\exp(2(T+i\theta)).$$

Irrespective of the choice of constants λ, H, and R, which are in general complex-valued, and functions $f(t)$ and $\mathcal{F}(t)$, expression (37) represents a solution to the homogeneous problem in question.

The values of H, λ, R, and ζ are chosen in such a way that the quantities

$$\xi_* = \zeta\exp(2(T+i\theta)); \quad \hat{y}_{bc} = \hat{y}_b + \hat{y}_c \qquad (38)$$

satisfy the relations

$$\mathcal{L}\hat{y}_{bc} + \mathcal{R}\xi_* = a^2\begin{bmatrix}\hat{A}_6\\\frac{1}{2}kE_0\end{bmatrix}\exp(2(T+i\theta)). \qquad (39)$$

Upon substituting (39) into (38), it appears that such values of λ, R, and ζ are obtained as solutions to a linear nonhomogeneous set of algebraic equations and are given by

$$\begin{bmatrix}H\\\lambda\\\zeta\\R\end{bmatrix} = \hat{L}^{-1}\left(\begin{bmatrix}C_4[2]\\2C_4[2]\rho\nu(k+q) - C_4[3]\\2iC_4[2]k - C_4[1](k+q)\\0\end{bmatrix} + \begin{bmatrix}\hat{A}_6\\\frac{1}{2}kE_0\end{bmatrix}\right). \qquad (40)$$

The expression for the square matrix $\hat{L}$ is given in Appendix C.

Since the columns $\hat{y}_a$ and $\hat{y}_{bc} = \hat{y}_b + \hat{y}_c$ identically satisfy relations (31), (32), and (39), the substitution

$$\hat{Y}_2 = \hat{y}_\alpha + \hat{y}_*; \quad \hat{y}_* = \hat{y}_a + \hat{y}_b + \hat{y}_c; \quad \xi = \xi_* \qquad (41)$$

in (27)–(30) leads to the problem

$$\mathcal{L}\hat{y}_\alpha = 0; \qquad (42)$$

$$\mathcal{B}\hat{y}_\alpha = a^2\left[\begin{bmatrix}\hat{A}_5\\\frac{1}{2}kE_0\end{bmatrix} - \begin{bmatrix}0\\f(t) + C_1[3] + C_2[3] + C_3[3]\\2C_1[1]q_1 + C_3[1](k+q)\\\mathcal{F}(t)\end{bmatrix}\right] \times \exp(2T); \qquad (43)$$

$$z \longrightarrow +\infty: \Phi_\alpha \equiv y_\alpha[4] \longrightarrow 0; \qquad (44)$$

$$z \longrightarrow -\infty: u_\alpha \equiv y_\alpha[1] \longrightarrow 0; \quad v_\alpha \equiv y_\alpha[2] \longrightarrow 0. \qquad (45)$$

If we set

$$f(t) = \hat{A}_5[2] - C_1[3] - C_2[3] - C_3[3];$$
$$\mathcal{F}(t) = \frac{1}{2}kE_0\exp(-2T), \qquad (46)$$

relations (43) are simplified to become

$$\mathcal{B}\hat{y}_\alpha = a^2\begin{bmatrix}0\\0\\A_5[3] - \Lambda\\0\end{bmatrix}; \qquad (43a)$$

$$\Lambda = 2C_1[1]q_1 + C_3[1](k+q).$$

We can now easily find (for details, see [18]) that

$$\hat{y}_\alpha = a^2\begin{bmatrix}A_5[3] - \Lambda\\0\\0\\0\end{bmatrix}\sqrt{\frac{\nu}{\pi}} \times \int_0^t \exp\left(-\frac{z^2}{4\nu(t-\tau)}\right)\exp(2S_1\tau)d\tau \qquad (47)$$

is the solution to the problem specified by Eq. (43)–(45) for the initial condition

$$t = 0: \hat{y}_\alpha = \hat{0}.$$

It can easily be verified that the substitution

$$\hat{Y}_2 = \hat{y}^* + \hat{y}_*, \qquad (48)$$

where

$$\hat{y}_* = \hat{y}_\alpha + \hat{y}_a + \hat{y}_b + \hat{y}_c;$$
$$\xi_2 = \xi^* + \zeta\exp(2(T+i\theta)), \qquad (49)$$

with $\hat{y}_\alpha$, $\hat{y}_a$ and $\hat{y}_b$, and $\hat{y}_c$ and ζ being calculated on the basis of, respectively, (47); (33)–(35); and (37), (40), and (46), reduces the second-order problem specified by Eqs. (27)–(30) to a homogeneous problem similar to that which is obtained in the first-order approxi-

mation; that is,

$$\mathcal{L}\hat{y}^* = \hat{0};$$

$$\mathcal{B}\hat{y}^* + \mathcal{R}\xi_2 = \hat{0};$$

$$z \longrightarrow +\infty: \Phi_2^* \equiv y^*[4] \longrightarrow 0;$$

$$z \longrightarrow -\infty: u_2^* \equiv y^*[1] \longrightarrow 0; \quad v_2^* \equiv y^*[2] \longrightarrow 0;$$

$$t = 0: \xi^* = \xi_2 - \zeta\exp(2ikx);$$

$$z \le 0: y^*[1]_{t=0} = u_2^0 - y_*[1]_{t=0};$$

$$y^*[2]_{t=0} = u_2^0 - y_*[2]_{t=0}.$$

Going over to the physical solution and considering that $\hat{y}_\alpha$ is constructed in such a way that, at the initial instant, all its elements are equal to zero, we can state that the real parts of the quantities forming the column $\hat{y}_*$ defined in (49) are the solutions to the second-order problem for the case where the initial conditions for these quantities are

$$t = 0: \xi_2 = \mathrm{Re}[\zeta\exp(2ikx)];$$

$$u_2^0 = \mathrm{Re}(y_a[1] + y_b[1] + y_c[1])_{t=0};$$

$$v_2^0 = \mathrm{Re}(y_a[2] + y_b[2] + y_c[2])_{t=0}.$$

The solution $\hat{y}^*$ obtained here does not become less significant because of the fact that it satisfies specific initial conditions. It is much more important that $\hat{y}^*$ is a key link of the substitution in (48), which reduces a nonhomogeneous second-order problem to a homogeneous one. The very fact that the new problem is homogeneous does not depend on the form of functions that specify the initial conditions for the second-order quantities. Therefore, the arbitrariness of formulation of initial conditions for these quantities affects only the initial conditions for the homogeneous problem to which the substitution in (48) reduces the original nonhomogeneous problem. But a homogeneous problem is the subject of investigations within well-developed linear wave theory.

PROFILE OF A NONLINEAR WAVE

If $\mathrm{Re}(\hat{y}_*)$ is added to the solution $\mathrm{Re}(\hat{y}^*)$ to the second-order problem, the resulting expression will again be a solution to the second-order problem. But the homogeneous problem of determining $\hat{y}^*$ coincides with the first-order problem specified by Eqs. (11)–(14). It follows that, if any solution of the form specified by (16)–(24) and characterized by different values of the wave number and the amplitude (the product of these quantities must be of higher order of smallness than ε) is added to $\mathrm{Re}(\hat{y}^*)$, we again arrive at a solution to the second-order problem. With allowance for the possibility of additions of this kind, the profile of the wave that runs along free charged surface of a deep viscous liquid can be described by the expression

$$\xi = \{a\cos(\theta)\exp(S_1 t) + a^2 A\cos(2\theta + \beta)\exp(2S_1 t)\} + \mathrm{LWS};$$

$$\zeta_1 = \mathrm{Re}(\zeta); \quad \zeta_2 = \mathrm{Im}(\zeta); \quad A = \sqrt{\zeta_1^2 + \zeta_2^2};$$

$$\beta = \begin{cases} \arctan\left(\dfrac{\zeta_2}{\zeta_1}\right). \text{ if } \quad \zeta_1 > 0 \text{ and } \zeta_2 < 0 \\ \arctan\left(\dfrac{\zeta_2}{\zeta_1}\right) + 2\pi, \quad \text{if} \quad \zeta_1 > 0 \text{ and } \zeta_2 < 0 \\ \arctan\left(\dfrac{\zeta_2}{\zeta_1}\right) + \pi, \quad \text{if} \quad \zeta_1 < 0 \\ \pi/2, \quad \text{if} \quad \zeta_1 = 0 \text{ and } \zeta_2 < 0 \\ 3\pi/2, \quad \text{if} \quad \zeta_1 = 0 \text{ and } \zeta_2 > 0, \end{cases} \tag{50}$$

where $\arctan(x)$ is that branch of the arctangent function which maps the numerical axis into the segment $(-\pi/2, \pi/2)$. The expression for the phase β is constructed in such a way that its value lies within the range $0 \le \beta < 2\pi$. The symbol ζ stands for a complex-valued quantity that, together with λ, H, and R, is determined by formula (40).

For the corrections considered in the preceding paragraph, we have adopted the notation LWS (linear wave solution), which is used in linear theory. This can be any superposition of running cosine waves $\eta\cos(\Sigma_2 t - \kappa x + \varphi)\exp(\Sigma_1 t)$ whose amplitude η and whose wave number $\kappa \ne k$ are such that $\kappa\eta = o(ka)$ (o is the symbol of smallness). The complex-valued frequency $\Sigma = \Sigma_1 + i\Sigma_2$ of these waves and their wave number κ are related by the same dispersion equation (21) as the quantities S and k in terms of which the variables $\theta = S_2 t - kx$ and $T = S_1 T$ are expressed. Being solutions to a homogeneous linear problem, LWSs propagate without interaction, each wave having its own phase velocity.

The terms in the braced expression on the right-hand side of (50) are interacting wave solutions. The coefficients A and β depend nontrivially on the wave number k corresponding to the wave $a\cos(\theta)$ propagating along the free surface in the first approximation. Thus, the amplitude and the phase of the second wave addition in the braced expression on the right-hand side of (50) depend on the properties of the first wave addition. Both wave components in the braces have the same phase velocity. Because of this, their sum is an individual periodic wave having the same phase velocity. For the length of this wave, it is natural to take the spatial period in x common to both terms—that is, $\lambda = 2\pi/k$. Obviously, k can be treated as the wave number corresponding to this new wave. In contrast to the wave solutions obtained by solving the problem in the linear

approximation, the profile of the new wave greatly depends on its wave number k and, as can be seen from (50), varies with time (in the case where $S_1 \neq 0$). Variations in the profile are of the second order of smallness.

As a result, we have found a solution to the problem of determining the profile of a nonlinear wave such that it involves two terms: a nonlinear wave proper whose profile depends on its wavelength and the LWS background that propagates according to the laws of linear wave theory. It is obvious that, in order to investigate nonlinear effects of the simulated phenomenon, it is not necessary to take into account the LWS noise in (50).

In the limit of a perfect liquid, in which case $\nu \longrightarrow 0$, β tends either to zero or to π. Noticing that, instead of $\beta = \pi$ we can set $\beta = 0$ and change the sign of A simultaneously, the profile of the wave in this approximation can be recast into a form that is instructive for investigations; that is,

$$\zeta = a\cos(\theta) + a^2 k\Lambda\cos(2\theta);\quad \Lambda = \frac{1}{4}\frac{(1+\alpha^2k^2-2\alpha kW)}{0.5-\alpha^2k^2};$$
$$W \equiv \frac{E_0^2}{4\pi\sqrt{\rho g\gamma}};\quad \alpha \equiv \sqrt{\frac{\gamma}{\rho g}}, \tag{51}$$

where α is the capillary constant and W is the dimensionless Tonks–Frenkel parameter characterizing the stability of the charged surface of a liquid with respect to the intrinsic charge. It is well known from linear theory that, at a preset value of αk, the condition

$$W > \alpha k + \frac{1}{\alpha k} \tag{52}$$

ensures the positivity of the parameter $S_1 \equiv \mathrm{Re}\,S$ appearing in (50). In this case, the free charged surface of a liquid is unstable [20]; the quantity S_1 here has the meaning of the increment of growth of the amplitude of the term that appears in the braced expression on the right-hand side of (50) and which is linear in a, while $2S_1$ has the meaning of the increment for the nonlinear term. If the value of W is not sufficient for the condition in (52) to be satisfied, then $S_1 < 0$. In this case, S_1 characterizes the decrement of an exponential decay of the term that appears in the braced expression on the right-hand side of (50) and which is linear in a, while $2S_1$ characterizes the decrement for the nonlinear term.

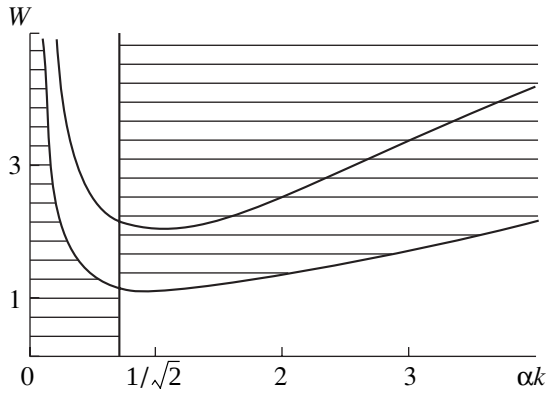


Fig. 1. Regions of the realization of various modes of a nonlinear wave motion. The region above the curve L corresponds to unstable solutions. In the shaded regions bounded by the curve Γ and the straight line $\alpha k = 1/\sqrt{2}$, waves have sharpened vertices, while, in the unshaded regions, waves have flattened vertices.

At $W = 0$, expression (50) coincides with the wave profile constructed in [9]. If, in addition to this, we set $\gamma = 0$, the well-known Stokes wave approximation quadratic in the amplitude will then be obtained. That nonlinear profile of a wave at the surface of a perfect liquid which is referred to as a Stokes wave was constructed more than a century ago. In the second-order approximation in the amplitude, expression (50) is nevertheless the first correct generalization of this concept to the case of a liquid that is characterized by a finite viscosity and a finite surface tension and which has a free charged surface. Thus, expression (50) itself is a new result for the theory of waves at the surface of a deep liquid and is a solution to a classic fluid dynamics problem that could not be solved for more than a century.

From (51), it follows that, in a perfect liquid at $k = k_* = (\alpha\sqrt{2})^{-1}$, the denominator of the amplitude of the second-order correction to the wave profile vanishes, while the correction itself tends to infinity at finite values of the numerator of Λ. This fact, which is associated with the nonlinear resonance interaction of waves, was partly discussed in [18]. In a viscous liquid, the amplitude of the wave remains finite [18], as may have been expected.

EFFECT OF AN ELECTRIC CHARGE ON THE PROFILES OF NONLINEAR WAVES

At a supercritical value of the surface charge density, electric forces have a more pronounced effect than Laplace forces, with the result that, at the charged surface, there appear outliers (Taylor cones), the charge being removed through their vertices via the emission of strongly charged drops [19, 21].

The condition in (52) specifies the region of unstable wave numbers (the region above the curve L in Fig. 1) in the plane of the dimensionless parameters αk and W. The value of $W_* = 2$, which corresponds to the value of $\alpha k = 1$ on the curve L, is the smallest value of W on the stability boundary. The wave $\zeta = a\cos(S_2 t - kx)\exp(T)$ characterized by the wave number $k = \alpha^{-1}$ becomes unstable at $W = W_*$. This wave is referred to as the main mode. The greater the value of W, the broader the spectrum of unstable sinusoidal wave solutions to the linearized problem. With increasing W, the spectrum of unstable waves broadens both toward low and toward high wave numbers. Within the concepts of linear theory, all unstable waves participate in the for-

mation of emission outliers. This means that, within linear theory, the bases of emission outliers must expand with increasing W, but this is not confirmed by a comparison of the experimental results reported in [21, 22], where the formation of emission outliers was observed at significantly different values of the field strength (different values of W). As a matter of fact, the mechanism through which long waves are involved in the process of formation of emission outliers obviously has some limitations that cannot be formulated within linear theory.

The nonlinear solution in (50) provides an alternative view of the formation of emission outliers. The main idealization employed by linear theory consists in the assumption that the superposition principle is valid. As a matter of fact, an emission outlier is formed not only owing to the summation of wave solutions to the linear problem but also owing to the growth of the amplitudes of nonlinear waves described by expressions similar to the first term in (50). It can easily be seen that, for $S_1 > 0$, the increment of growth of the amplitude of the term that is nonlinear in a with increasing time is twice as great as that for the term that is linear in the amplitude. Within a short time interval, it is not the solution to the linear problem but the correction to it corresponding to a nonlinear character of the wave that plays a decisive role in the formation of outliers.

The solution in (50) can be considered as a model one in which, at the initial instant of time, the wave profile is described in the first approximation by only one harmonic solving the linear problem. In the general case where the initial profile is determined by a superposition of a number of such harmonics, the solution will become much more complicated; instead of the second term in the braced expression on the right-hand side of (50), there then arises the sum of all possible waves generated by the three-mode wave interaction [23]. Their increments are determined by pair products of the increments of original waves. It follows that the rate of their growth is higher than that for the first-approximation waves and that the formation of emission outliers is determined by waves generated by a nonlinear interaction.

Yet another interesting property of a wave having the profile described by the braced expression on the right-hand side of (50) manifests itself if, in the $(\alpha k, W)$ plane, where we have already constructed the stability boundary for solutions to the linear problem, we depict one more curve Γ, that which is determined by the equation

$$\frac{1}{2}\left(\alpha k + \frac{1}{\alpha k}\right) - W = 0. \qquad (53)$$

From (51), it follows that the amplitude of the second order of smallness vanishes on this curve, $\Lambda = 0$; to third-order terms inclusive, the profile of a periodic wave for $(k, W) \in \Gamma$ in the limit of a perfect liquid is exactly sinusoidal (without any correction). Calculations by formula (50) reveal that, in the case of a viscous liquid, the quantities A as functions of the parameters αk and W attain a local minimum on this curve (they are close to zero at this minimum).

The next point that deserves particular attention is associated with the role of the phase β appearing in (50). From (50) and (51), it follows that, at $A \neq 0$, a wave that, owing to the fact that $a\cos(\theta)$ and the short-wavelength nonlinear correction $A\cos(2\theta + \beta)$ are cophased, has a greater curvature of the vertex than the first-approximation wave (in the nonlinear approximation, the vertex of the wave becomes sharper) corresponds to a phase value close to $\beta = 0$. At $\beta = \pi$ and $A \neq 0$, the vertex takes on the contrary a more rounded shape.

In analyzing the expression for the amplitude of the nonlinear correction Λ, one can easily see that $\Lambda > 0$ if

$$\begin{cases} k^2 < 0.5 \\ W < 0.5(\alpha k + (\alpha k)^{-1}) \end{cases}$$

or

$$\begin{cases} k^2 > 0.5 \\ W > 0.5(\alpha k + (\alpha k)^{-1}) \end{cases}$$

and $\Lambda < 0$ if

$$\begin{cases} k^2 < 0.5 \\ W > 0.5(\alpha k + (\alpha k)^{-1}) \end{cases}$$

or

$$\begin{cases} k^2 > 0.5 \\ W < 0.5(\alpha k + (\alpha k)^{-1}). \end{cases}$$

This means that the intersecting curve Γ and straight line $\alpha k = 1/\sqrt{2}$ divide the plane of the parameters k and W into four regions restricted by the angles, each having a straight-line and a curvilinear side (see Fig. 1). In the limit $\nu \longrightarrow 0$, the pairs of opposite angles bound regions where the phase β takes identical values. Within the shaded regions, the phase vanishes, $\beta = 0$, and nonlinear waves have sharp vertices (see Figs. 2a, 2c). The phase value of $\beta = \pi$ corresponds to the other pair of regions where the vertices of nonlinear waves are smoothed (Figs. 2b, 2d).

In the limit $\nu \longrightarrow 0$, the phase β undergoes jumplike changes on the sides of the angles. If the viscosity is taken into account, the sides of the angles in the $(\alpha k, W)$ plane are smeared into bands where the phase β changes smoothly.

From Fig. 1, it can be seen that the region where the solutions to the linearized problem are unstable (the region above the curve L) is divided into two parts by

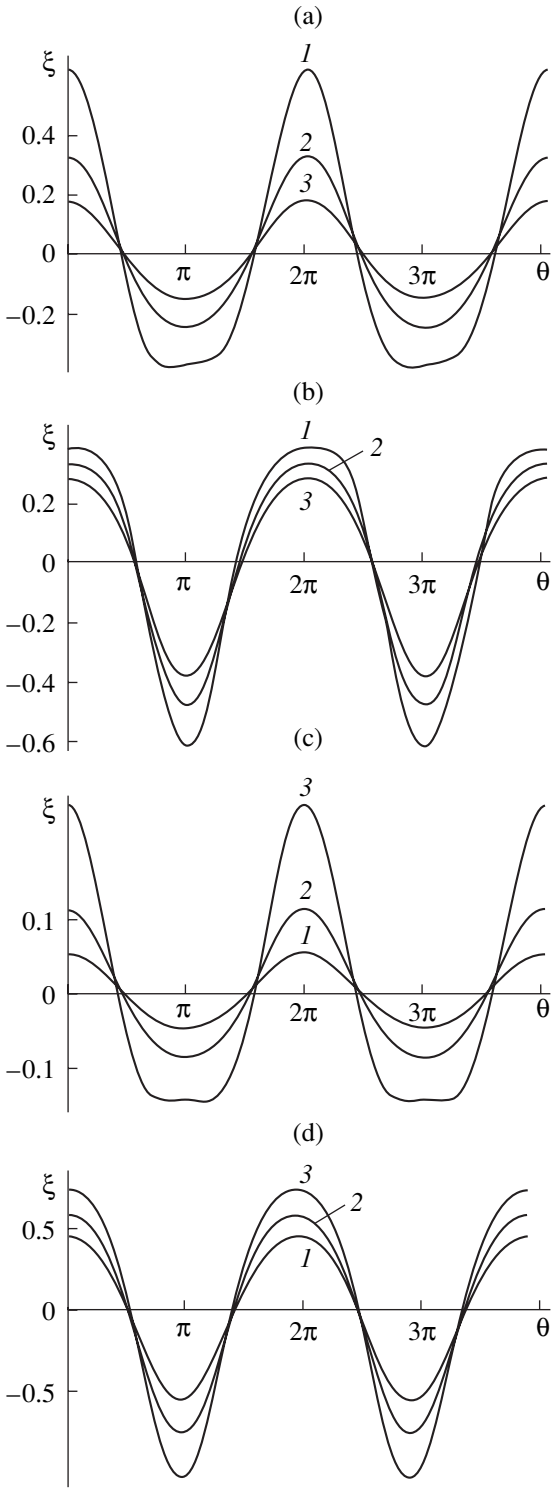


Fig. 2. Profiles of nonlinear waves of different types in dimensionless variables where $\rho = g = \gamma = 1$ for $\nu = 10^{-2}$ at the instants of time t = (*1*) 0, (*2*) 30, and (*3*) 60: (a, b) profiles of the waves whose amplitudes decrease with time because of viscous damping for the Tonks–Frenkel parameter taking the subcritical value of $W = 1.5$ [k = (a) 1, (b) 0.6] and (c, d) profiles of the waves whose amplitudes increase with time owing to the realization of instability with respect to the surface charge for $W = 2.5$ [k = (c) 1, (d) 0.6].

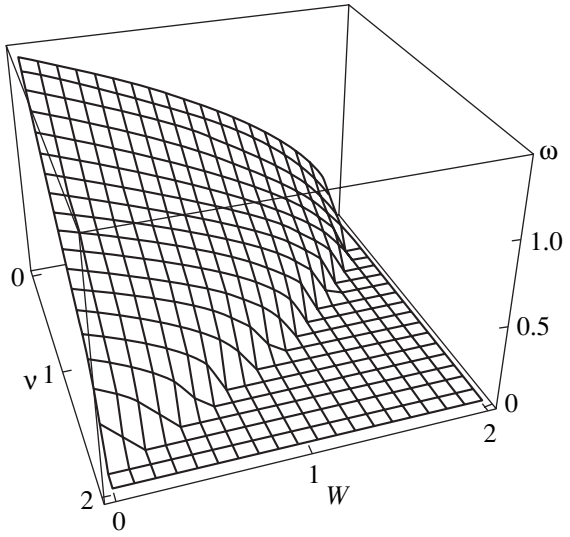


Fig. 3. Dimensionless frequency of wave motion as a function of the dimensionless viscosity ν and the dimensionless parameter W according to calculations at $k = 1$.

the straight line $\alpha k = 1/\sqrt{2}$. For a wave of the type in (50), the development of instability is accompanied by the formation of a sharp point at the vertex if $\alpha k > 1/\sqrt{2}$ (see Fig. 2c) and by the flattening of the vertex if $\alpha k < 1/\sqrt{2}$ (see Fig. 2d). At the vertices of the outliers, the electric field tears off drops of characteristic linear dimension about the radius of curvature of the vertex [18, 21]. Therefore, the emission of drops from a sharpened vertex due to the separation of smaller and more strongly charged drops is more probable than the separation of drops from the flattened vertex of a wave. After the commencement of emission, the time evolution of waves follows new regularities that must be explored individually. Thus, an analysis of the nonlinear solution to the problem leads to a natural limitation from the side of long waves at wavelengths that are involved in the formation of emission outliers.

In the second-order approximation used here, neither viscosity nor nonlinearity affects the conditions for the realization of instability of the charged surface of a liquid. Their effect will manifest itself in the calculations of the next order of smallness as corrections to the frequencies of the waves (we recall that the critical conditions for the realization of instability of the charged surface of a liquid are determined by the vanishing of the frequency squared [19]).

The surface-charge and viscosity effect on the frequencies of the waves that follows from a linear analysis is illustrated in Fig. 3.

CONCLUSIONS

An analysis of the solution obtained here for the problem of nonlinear periodic waves propagating along

the charged surface of a conducting viscous liquid has revealed that the main contribution to the process of formation of emission outliers that arise upon the realization of instability of the charged surface comes not from the entire set of instabilities of waves whose wavelengths take values in the interval predicted by linear theory but only from that part of them for which the nonlinearity of the process manifests itself in the sharpening of the vertices. The wave numbers of such waves satisfy the condition $k > k_* = \sqrt{\rho g/(2\gamma)}$.

APPENDIX A

At each point $z = \xi(x, z, t)$ of the flat free liquid surface distorted by wave motion, we define the following quantities:

(i) the unit vectors $\mathbf{n}$ and $\boldsymbol{\tau}$ along, respectively, the external normal and the tangent,

$$\mathbf{n} = \frac{\nabla(z-\xi)}{|\nabla(z-\xi)|} = \frac{-\mathbf{e}_x\partial_x\xi + \mathbf{e}_z}{\sqrt{1+(\partial_x\xi)^2}}; \quad \boldsymbol{\tau} = \frac{\mathbf{e}_x + \mathbf{e}_z\partial_x\xi}{\sqrt{1+(\partial_x\xi)^2}};$$

(ii) the mean curvature of the surface,

$$\mathrm{div}(\mathbf{n}) = -\frac{\partial_{xx}\xi}{(1+(\partial_x\xi)^2)^{3/2}} = -\partial_{xx}\xi + O(\xi^3).$$

APPENDIX B

(i) In the generally adopted notation, the kinematic boundary conditions for quantities of different order of smallness can be derived from (4) [18] and are given by

$$z = 0: \partial_t\xi_1 - v_1 = 0; \quad \partial_t\xi_2 - v_2 = \xi_1\partial_z v_1 - u_1\partial_x\xi_1.$$

(ii) For the quantities of different order of smallness, the conditions for the tangential tensions at a free surface can be derived from (6) [18] and are given by

$$z = 0: \partial_x v_1 + \partial_z u_1 = 0;$$

$$\partial_x v_2 + \partial_z u_2 = -4\partial_z v_1\partial_x\xi_1 - \xi_1\partial_z(\partial_x v_1 + \partial_z u_1).$$

(iii) The condition in (5) for the pressure at the disturbed surface differs from that considered in [18] by the presence of the term $(8\pi)^{-1}(\nabla\Phi)^2$ on the right-hand side. For this term, the expansion in ξ in the vicinity of $z = 0$ has the form

$$z = \xi: \frac{1}{8\pi}((\nabla\Phi)^2 + 2\xi(\nabla\Phi\cdot\partial_z(\nabla\Phi))$$

$$+\xi^2(\partial_z(\nabla\Phi))^2 + \xi^2(\nabla\Phi\cdot\partial_{zz}(\nabla\Phi))_{z=0} + O(\varepsilon^3).$$

If we consider that $\xi = \xi_1 + \xi_2 + O(\varepsilon^3)$, where $\xi_1 \sim O(\varepsilon)$ and $\xi_2 \sim O(\varepsilon^2)$, $\Phi = \Phi_0 + \Phi_1 + \Phi_2 + O(\varepsilon^3)$, where $\Phi_1 \sim O(\varepsilon)$ and $\Phi_2 \sim O(\varepsilon^2)$, and $\nabla\Phi_0 = -E_0\mathbf{e}_z$ and $\partial_z(\nabla\Phi_0) = \mathbf{O}$, relations similar to those in [18] can easily be derived for quantities of different order from the condition in (5) for the pressure at a free surface. Owing to the appearance of the electric-field pressure in the present case, they become

$$z = 0: p_0 = \frac{E_0^2}{4\pi};$$

$$-\rho g\xi_1 + p_1 - 2\rho\nu\partial_z v_1 + \gamma\partial_{xx}\xi_1 - \frac{1}{4\pi}E_0\partial_z\Phi_1 = 0;$$

$$-\rho g\xi_2 + p_2 - 2\rho\nu\partial_z v_2 + \gamma\partial_{xx}\xi_2$$

$$-\frac{1}{4\pi}E_0\partial_z\Phi_1 = 2\rho\nu\xi_1\partial_{zz}v_1 - \xi_1\partial_z p_1$$

$$-\frac{1}{8\pi}((\partial_x\Phi_1)^2 + (\partial_z\Phi_1)^2) + \frac{1}{4\pi}E_0\xi_1\partial_{zz}\Phi_1.$$

(iv) To the required degree of precision, the expansion of the condition in (7) in the vicinity of the undisturbed surface of a liquid has the form

$$z = 0: \Phi + \xi\partial_z\Phi + \frac{1}{2}\xi^2\partial_{zz}\Phi = 0.$$

Upon substituting $\xi = \xi_1 + \xi_2 + O(\varepsilon^3)$, where $\xi_1 \sim O(\varepsilon)$ and $\xi_2 \sim O(\varepsilon^2)$, $\Phi = \Phi_0 + \Phi_1 + \Phi_2 + O(\varepsilon^3)$, where $\Phi_1 \sim O(\varepsilon)$, $\Phi_2 \sim O(\varepsilon^2)$, and $\Phi_0 = -E_0 z$, into this expansion, we can easily obtain the following relations for quantities of different order of smallness:

$$z = 0: \Phi_0 = 0; \ \Phi_1 - E_0\xi_1 = 0; \ \Phi_2 - E_0\xi_2 = -\xi_1\partial_z\Phi_1.$$

APPENDIX C

The quantities S, S_1, S_2, q, q_1, q_2, and D are calculated with the aid of expressions (23) and (24).

(i) The matrix-columns $\hat{A}_1$, $\hat{A}_2$, $\hat{A}_3$, $\hat{A}_4$, $\hat{A}_5$, and $\hat{A}_6$ appearing in (27) and (28) are given by

$$\hat{A}_1 = -4\nu^2k^3q_1\begin{bmatrix} q_2 \\ k \\ 0 \end{bmatrix}; \quad \hat{A}_2 = \begin{bmatrix} 0 \\ -k(S_2^2 + D^2) \\ 0 \end{bmatrix};$$

$$\hat{A}_3 = \nu k\begin{bmatrix} 2(\bar{S} + 2\nu k^2)q_1q_2 + (S_2 + iD)(q_2^2 - q_1^2 + k^2) \\ -2ik(k+q)(S_2 + iD) \\ 0 \end{bmatrix};$$

$$\hat{A}_4 = \begin{bmatrix} -i\nu k(q-k)^2(S + 2\nu k^2) \\ 0 \\ 0 \end{bmatrix};$$

$$\hat{A}_5 = \begin{bmatrix} 0 \\ N_0 \\ M_0 \end{bmatrix}; \quad \hat{A}_6 = \begin{bmatrix} \Omega \\ N \\ M \end{bmatrix}; \quad \begin{aligned} \Omega &= \Omega_1 + i\Omega_2; \\ N &= N_1 + iN_2; \\ M &= M_1 + iM_2, \end{aligned}$$

where $\bar{S}$ is the complex conjugate of S and N_1, N_2, Ω_1, Ω_2, M_1, and M_2 are calculated by the formulas

$$\Omega_1 = k(D - 2\nu k q_1); \quad \Omega_2 = k(S_2 - 2\nu k q_2);$$

$$N_0 = \frac{1}{2}\rho(S_2^2 - S_1^2 + 4\nu k^2 S_1 + 4\nu^2 k^2 (k^2 - q_1^2 + q_2^2));$$

$$N_1 = N_0 + \frac{1}{2}k^2 \frac{E_0^2}{4\pi}; \quad N_2 = \rho(S_2 D - 4\nu^2 k^2 q_1 q_2);$$

$$M_0 = k(kS_2 - q_2(3k^2 + q_2^2 - 3q_1^2));$$

$$M_1 = -k(3kS_2 - q_2(5k^2 - q_2^2 + 3q_1^2)\nu);$$

$$M_2 = k(3kS_1 + \nu(6k^3 - 5k^2 q_1 + 3q_2^2 q_1 - q_1^3)).$$

(ii) The quadratic matrices $\hat{\Pi}_1$, $\hat{\Pi}_2$, $\hat{\Pi}_3$, and $\hat{\Pi}_4$ appearing in (35) can be written as

$$\hat{\Pi}_1 = \begin{bmatrix} 2(S_1 - 2\nu q_1^2) & 0 & 0 \\ 0 & 0 & 2q_1\rho^{-1} \\ 0 & 2q_1 & 0 \end{bmatrix};$$

$$\hat{\Pi}_2 = \begin{bmatrix} 2(S_1 - 2\nu k^2) & 0 & 0 \\ 0 & 0 & 2k\rho^{-1} \\ 0 & 2k & 0 \end{bmatrix};$$

$$\hat{\Pi}_3 = \begin{bmatrix} 2S_1 - \nu(k+q)^2 & 0 & 0 \\ 0 & 0 & (k+q)\rho^{-1} \\ 0 & k+q & 0 \end{bmatrix};$$

$$\hat{\Pi}_4 = \begin{bmatrix} 2S + \nu(3k^2 - 2kq - q^2) & 0 & -2ik\rho^{-1} \\ 0 & 2s + \nu(3k^2 - 2kq - q^2) & (k+q)\rho^{-1} \\ -2ik & k+q & 0 \end{bmatrix}.$$

(iii) The matrix $\hat{L}$ appearing in (40) has the form

$$\hat{L} = \begin{bmatrix} 2ik & -2k & 2S & 0 \\ 4i\rho\nu kr & -2\rho(S + 4\nu k^2) & -(\rho g + 4\gamma k^2) & kE_0/(2\pi) \\ 8k^2 + 2S/\nu & -8ik^2 & 0 & 0 \\ 0 & 0 & -E_0 & 1 \end{bmatrix}.$$

REFERENCES

1. G. Lamb, *Hydrodynamics* (Cambridge Univ. Press, Cambridge, 1932; Gostekhizdat, Leningrad, 1947).
2. J. H. Michel, Philos. Mag. **36**, 430 (1893).
3. J. R. Wilton, Philos. Mag. **27**, 385 (1914).
4. J. R. Wilton, Philos. Mag. **29**, 385 (1915).
5. W. J. Harrison, Proc. London Math. Soc. **7**, 107 (1908).
6. J. J. Stoker, *Water Waves* (Wiley, New York, 1957; Inostrannaya Literatura, Moscow, 1959).
7. G. B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974; Mir, Moscow, 1977).
8. R. V. Gol'dshteĭn and V. A. Gorodtsov, *Mechanics of Continuous Media* (Nauka, Moscow, 2000), Part 1.
9. A. H. Nayfeh, J. Fluid Mech. **48**, 385 (1971).
10. I. Shugan and K. Voliak, J. Fluid Mech. **368**, 321 (1998).
11. K. Trulsen and I. Kliakhandler, Phys. Fluids **12**, 2432 (2000).
12. M. S. Longet-Higgins, J. Fluid Mech. **423**, 275 (2000).
13. S. O. Shiryaeva, D. F. Belonozhko, V. B. Svetovoĭ, and A. I. Grigor'ev, Preprint No. 31 (Institute of Microelectronics and Informatics, Russian Academy of Sciences, Yaroslavl, 2001).
14. I. M. Mindlin, Izv. Ross. Akad. Nauk, Mekh. Zhidk. Gaza, No. 3, 135 (1994).
15. S. V. Nesterov, Izv. Ross. Akad. Nauk, Mekh. Zhidk. Gaza, No. 4, 116 (1995).
16. A. V. Fedorov and W. K. Melvil, J. Fluid Mech. **354**, 1 (1998).
17. D. F. Belonozhko, A. I. Grigor'ev, and S. O. Shiryaeva, Pis'ma Zh. Tekh. Fiz. **28** (19), 1 (2002) [Tech. Phys. Lett. **28**, 795 (2002)].
18. D. F. Belonozhko and A. I. Grigor'ev, Zh. Tekh. Fiz. **73** (4), 28 (2003) [Tech. Phys. **48**, 404 (2003)].
19. A. I. Grigor'ev and S. O. Shiryaeva, Izv. Ross. Akad. Nauk, Mekh. Zhidk. Gaza, No. 3, 3 (1994).
20. Ya. I. Frenkel, Zh. Éksp. Teor. Fiz. **6**, 348 (1936).
21. M. D. Gabovich, Usp. Fiz. Nauk **140**, 137 (1983) [Sov. Phys. Usp. **26**, 447 (1983)].
22. M. B. Voĭtsekhovskiĭ, Dokl. Akad. Nauk SSSR **262** (1), 87 (1982) [Sov. Phys. Dokl. **27**, 54 (1982)].
23. O. M. Philips, J. Fluid Mech. **106**, 215 (1981).

Translated by A. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1407–1412. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 47–52.

GASES AND LIQUIDS

Simplified Description of Small-Scale Turbulence

A. M. Balonishnikov

St. Petersburg State University of Engineering and Economics, St. Petersburg, 197002 Russia
e-mail: balonalex@yahoo.co.uk

Received April 2, 2003

Abstract—A simplified description of small-scale anisotropic turbulence is developed in the form of a set of three integrodifferential equations in one-dimensional momentum space or in the form of a set of three partial differential equations in a modified one-dimensional physical space. In the first case, the unknown functions are three coefficients in the Taylor expansion of an unstable polarization Fourier harmonic of the pulsating component of the velocity near the most unstable direction, the independent variables being time and the absolute value of the wave vector. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The problem of reducing the Navier–Stokes equations that describe well-developed turbulence to a set of equations with a smaller number of degrees of freedom is important not only from a theoretical standpoint but also for practical calculations of turbulent flows in various devices. Small-scale turbulent processes (from energetic to dissipative vortices) possess an especially large number of degrees of freedom. Modern physical approaches to describing small-scale turbulence can be found in [1, 2].

One possible approach to constructing the models of large-scale turbulence in an incompressible fluid is to separate the velocity and pressure fields into large- and small-scale components (pulsations). If we could solve the equations for small-scale components at least approximately, then, by substituting the solutions obtained into the equations for large-scale turbulent components, we would arrive at equations exclusively for large-scale components. This approach has so far been implemented in the linear approximation only by using the concept of "random force" [3, 4], whose parameters are to be specified in advance.

Although the approach proposed here does not allow the pulsating components to be expressed through large-scale fields, this is, to the best of the author's knowledge, the first approach that makes it possible to reduce the three-dimensional problem of the turbulence closure to a quasi-one-dimensional problem.

DERIVATION OF THE BASIC EQUATIONS IN MOMENTUM SPACE

We start with the following basic equation for Fourier harmonics of the polarization components of the velocity [5]:

$$(\partial_t + \nu k^2) v^\lambda = J^{\lambda\alpha} v^\alpha + (\mathbf{B}^{-1})^{\lambda\gamma} \Sigma_{\alpha,\beta,\mathbf{p}+\mathbf{q}=\mathbf{k}} \Phi^{\alpha\beta\gamma}(\mathbf{Bv})^\alpha(\mathbf{p}, t)(\mathbf{Bv})^\beta(\mathbf{q}, t). \quad (1)$$

Here, $\mathbf{k}$, $\mathbf{p}$, and $\mathbf{q}$ are the wave vectors whose three Cartesian coordinates take on the values k_i, p_i, $q_i = \pm 2\pi n/L$ (where $i = 1, 2, 3$ and $n = 1, 2, 3, \ldots$); L is the scale separating large- and small-scale motions; the upper polarization indices α, β, γ, and λ take on the values 1 and 2; ν is the kinematic molecular viscosity coefficient; $\mathbf{v}$ is the two-dimensional polarization velocity vector, $\mathbf{v}(k, \theta, \eta, t)$, that is sought in the spherical coordinate system used in [6], in which case the wave vector has the form $\mathbf{k} = (k\cos\theta\cos\eta, k\sin\theta\cos\eta, k\sin\eta)$; $\mathbf{J}$ is a diagonal matrix; and $\mathbf{B}$ is the matrix of transformation of the polarization components of the small-scale velocity.

The diagonal elements of the matrix $\mathbf{J}$ are the eigenvalues $\lambda_{1,2}$ of the matrix $\mathbf{A}$ of the linear (without the viscous terms) basic equations [5]:

$$\lambda_{1,2} = \mathbf{n}\cdot\mathbf{Sn}/2 \pm \sqrt{(\mathbf{n}\cdot\mathbf{Sn})^2/4 + \mathrm{tr}[(\mathbf{k}\times\mathbf{S})^2] - (\mathbf{n}\cdot\Omega)^2/4}. \quad (2)$$

Here, $\mathbf{n} = \mathbf{k}/k$; the large-scale velocity gradient tensor $\mathbf{S}$ with the elements $S_{ij} = \frac{1}{2}(\partial_i U_j + \partial_j U_i)$; U_j (with $j = 1, 2, 3$) are the components of the large-scale velocity vector; $\partial_j U_i$ is the partial derivative of the ith velocity component with respect to the jth Cartesian spatial coordinate; tr is the trace of the matrix; and $\mathbf{\Omega}$ is the large-scale vorticity vector, $\mathbf{\Omega} = \nabla \times \mathbf{U}$.

The elements of the matrix $\mathbf{A}$ are defined by the relationship $A^{\gamma\mu} = -\varepsilon_j^\gamma \varepsilon_m^\mu \partial_m U_j$ with summations over repeated indices. In this relationship, $\boldsymbol{\varepsilon}^1$ and $\boldsymbol{\varepsilon}^2$ are two unit polarization vectors that are perpendicular to the wave vector $\mathbf{k}$ and are defined as $\boldsymbol{\varepsilon}^1 = (\sin\theta, -\cos\theta, 0)$

1063-7842/03/4811-1407$24.00

and $\boldsymbol{\varepsilon}^2 = (\cos\theta\sin\theta, \sin\theta\sin\eta, -\cos\eta)$, satisfying the equalities $\boldsymbol{\varepsilon}^1(-\mathbf{k}) = -\boldsymbol{\varepsilon}^{-1}(\mathbf{k})$ and $\boldsymbol{\varepsilon}^2(-\mathbf{k}) = \boldsymbol{\varepsilon}^2(\mathbf{k})$, which yield the relationships $v^1(-\mathbf{k}) = -v^{1*}(\mathbf{k})$ and $v^2(-\mathbf{k}) = v^{2*}(\mathbf{k})$ by virtue of the fact that the velocity is a real quantity. The rest of the notation in Eq. (1) is

$$\Phi^{\gamma\alpha\beta} = -ik_m\varepsilon_j^\gamma(\mathbf{k})\varepsilon_j^\alpha(\mathbf{p})\varepsilon_m^\beta(\mathbf{q}).$$

In the approach proposed here, the large-scale velocity is characterized by a constant component, which is eliminated by a Galilean transformation, and a constant velocity gradient tensor $\mathbf{dU}$ in considering the dynamics of the small-scale velocity. This is in contrast to the approach developed in [4], in which the large-scale velocity is assumed to be linear. That is why these two approaches yield different equations for the small-scale velocity (see [5] for details).

According to the Haken slaving principle [8], the unstable modes, which are the order parameters of the problem, predominate over the stable modes. We can even go beyond the Haken principle and assume that the main contribution to the energetics of the process comes from the most unstable modes, which thus should be described with the highest accuracy. In [8], it was also assumed that the suppression of unstable modes should be described by cubic terms and that the most unstable modes are singled out by a kind of Laplacian operator. In comparison with this approach, the method proposed in the present paper is simpler: it consists in expanding both dependent and independent variables in Taylor series in two angular variables about the points corresponding to the fastest growing Fourier harmonics of the velocity.

In order to illustrate the method, we consider a Couette flow such that the large-scale velocity gradient tensor has the only nonzero component $S_{12} = \partial_1 U_2$. In this case, we have $\lambda_1 = 1/2\sin(2\theta)\cos^2\eta$ and $\lambda_2 = 0$. The quantity λ_1 takes on the same maximum value at the two points ($\theta_1 = \pi/4$, $\eta_1 = 0$) and ($\theta_2 = 5\pi/4$, $\eta_2 = 0$). Retaining several first terms in the Taylor series expansions of the function λ_1 near these points, we arrive at the expansion

$$\lambda_1 = \frac{S}{2}[1 - 2(\theta - \theta_1)^2 - (\eta - \eta_1)^2], \qquad (3)$$

with $i = 1, 2$.

Accordingly, the expansions of the elements of the matrices $\mathbf{B}$ and $\mathbf{B}^{-1}$, which are composed of the eigenvectors of the original matrix $\mathbf{A}$, have the form $b_{11} = 1$ and $b_{22} = 1$ (the matrix $\mathbf{B}$ can be chosen with a certain amount of freedom because the eigenvectors are defined to within a numerical factor) and also $b_{21} = -(\eta - \eta^3/6)[1 + 2(\theta - \theta_1)]$ and $b_{21} = -(\eta - \eta^3/6)[1 - 2(\theta - \theta_1)]$.

The elements of the inverse matrix $\mathbf{B}^{-1}$ are equal to

$$b_{11}^{-1} = 1 + \eta^2, \quad b_{22}^{-1} = 1 + \eta^2,$$

$$b_{12}^{-1} = \left(\eta + \frac{7}{6}\eta^3\right)[1 - 2(\theta - \theta_i)],$$

$$b_{21}^{-1} = \left(\eta + \frac{7}{6}\eta^3\right)[1 + 2(\theta - \theta_i)].$$

The sum in the basic equation (1) for the polarization Fourier harmonics of the velocity $\mathbf{v}$, which are modified by the matrix $\mathbf{B}$, contains the wave vectors $\mathbf{k}$, $\mathbf{p}$, and $\mathbf{q}$ (such that $\mathbf{q} = \mathbf{k} - \mathbf{p}$), whose components in spherical coordinates are (k, θ_k, η_k), (p, θ_p, η_p), and (q, θ_q, η_q), respectively.

It is easy to show that, in the spherical coordinate system adopted here and in [7], the following relationships are satisfied:

$$\tan\theta_q = \frac{k\sin\theta_k\sin\eta_k - p\sin\theta_p\cos\eta_p}{k\cos\theta_k\cos\eta_k - p\cos\theta_p\cos\eta_p}, \qquad (4)$$

$$\sin\eta_q = (k\sin\eta_k - p\sin\eta_p)/q, \qquad (5)$$

where

$$q = \sqrt{k^2 + p^2 - 2kp\cos\eta_k\cos\eta_p\cos(\theta_k - \theta_p) - 2kp\sin\eta_k\sin\eta_p}.$$

In this case, the expansions of θ_q and η_q about the directions determined by the angles θ_i and η_i have the form

$$\theta_q = \theta_i + \frac{k}{k-p}(\theta_k - \theta_i) + \frac{p}{p-k}(\theta_p - \theta_i),$$

$$\eta_q = \frac{l}{|k-p|}\eta_k - \frac{p}{|k-p|}\eta_p.$$

From physical considerations, it is clear that $q = |k - p| \le 2\pi\sqrt{3}/L$ and $p \le 2\pi\sqrt{3}/L$, where the scale L separates small-scale pulsations from large-scale motions.

In numerical modeling involving large-scale vortices, the scale L can be thought of as the spacing of the numerical grid. The form of the above terms in the Taylor series expansions and the subsequent terms, which have been discarded, leads us to the conclusion that the approximation proposed here is poor for large values of the ratio $k/(k - p)$; moreover, it is very poor for $kL \gg 1$, i.e., for small-scale vortices whose sizes are much less than the grid spacing. Fortunately, from experiments, it is known that, in the range of scales in question (such that the grid spacing L is smaller than the global scale of turbulence), the spectra of turbulent Reynolds stresses and the energy spectra in the inertial interval decrease according to power laws (the power indices being close to –5/3 and –2, respectively), and, for dissipative vortices, the spectra decrease exponentially [1, 2]. Consequently, when the contribution of solutions to the model equations to the spectra of the energy and Reynolds stresses are not too large, the above poor approximation does not significantly worsen the

description of the dynamics of the largest of small-scale vortices, which make the main contribution to subgrid Reynolds stresses.

Following the Haken subordination principle [8], we can set $v^2 = 0$ for the stable and neutrally stable modes under consideration. Near the most unstable directions determined by the angles θ_i and η_i, the largest contribution comes from nonlinear terms with the coefficients

$$\Phi^{111} \approx ik\frac{p}{p-k}(\theta_k - \theta_p),$$

$$\Phi_{112} \approx -ik\left(\frac{k}{|k-p|}\eta_k - \frac{p}{|k-p|}\eta_p - \eta_k\right),$$

the remaining coefficients $\Phi^{\alpha,\beta,\gamma}$ being polynomials of higher orders in the angular variables. As in [7], in the limit $L \longrightarrow \infty$, we can switch from summation to integration over the wavenumbers in Eq. (1):

$$(2\pi/L)^3\Sigma_k \longrightarrow \int d^3k.$$

Using the above Taylor series expansions, we obtain the following approximate equation for the first velocity component v^1 of unstable modes:

$$(\partial_t + \nu k^2)v^1 = \lambda_1 v^1 + ik\left(\frac{L}{2\pi}\right)^3 \int_{2\pi/L}^{\infty} p^2 dp \times \int_{-\pi/2}^{\pi/2} \cos\eta_p d\eta_p \int_{\theta'_p} d\theta_p Z v^1(\mathbf{p}) v^1(\mathbf{q}), \quad (6)$$

where θ'_p are the angles corresponding to the unstable modes,

$$\theta'_p \in \left[0, \frac{\pi}{2}\right] \cup \left[\pi, \frac{3\pi}{2}\right],$$

and

$$\lambda_1 = \frac{S}{2}[1 - 2(\theta - \theta_j)^2 - (\eta - \eta_j)^2]; \quad j = 1, 2;$$

$$Z = -\frac{p}{k-p}(\theta_k - \theta_j) - \frac{p}{k-p}(\theta_p - \theta_j) + \left(\frac{k}{|k-p|}\eta_k - \frac{p}{|k-p|}\eta_p - \eta_k\right)\left(\frac{k}{|k-p|}\eta_k - \frac{p}{|k-p|}\eta_p\right).$$

The spherical coordinate system [7] that was conveniently used up to this point can be changed into a "hemispherical" coordinate system in which the components of the wave vector are related to those in the Cartesian coordinate system in the above manner, $\mathbf{k} = (k\cos\theta\cos\eta, k\sin\theta\cos\eta, k\sin\eta)$, but, in the lower half-plane, the quantity k is negative, the angle η spans the same range,

$$\eta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right],$$

and the range spanned by the angle θ is two times shorter, $\theta \in [0, \pi]$. In this hemispherical coordinate system, the function λ_1 has a maximum only at one point, namely, at (θ_1, η_1). In this case, the integration limits in Eq. (6) change as follows:

$$(\partial_t + \nu k^2)v^1 = \lambda_1 v^1 + ik\left(\frac{L}{2\pi}\right)^3 \int_{-\infty}^{\infty} p^2 dp \times \int_{-\pi/2}^{\pi/2} \cos\eta d\eta \int_0^{\pi/2} d\theta Z v^1(\mathbf{p}) v^1(\mathbf{q}). \quad (7)$$

Since we seek the small-scale velocity, we assume that the large-scale velocity components are identically zero:

$$v^1(\mathbf{p}) = 0, \quad |\mathbf{p}| < \frac{2\pi}{L}; \quad v^1(q) = 0, \quad |\mathbf{q}| < \frac{2\pi}{L};$$

$$v^1(\mathbf{k}) = 0, \quad |\mathbf{k}| < \frac{2\pi}{L}.$$

The method for solving an integral equation by expanding its kernel and the unknown function in Taylor series about a certain point (see, e.g., [9]) is known but has not become widely used. Here, this method is applied to an integrodifferential equation and with respect to only some of the variables (in the case at hand, the angular variables). Assuming that the velocity v^1 has a maximum with respect to the angular variables at the same point at which λ_1 is maximum, namely, at (θ_1, η_1), we arrive at the following expansions:

$$v^1(k, \theta_k, \eta_k, t) = v_0(k, t) + c(k, t)(\theta_k - \theta_1)^2 + d(k, t)\eta_k^2 + f(k, t)\eta_k(\theta_k - \theta_1) \quad (8)$$

and, accordingly,

$$v^1(q, \theta_q, \eta_q, t) = v_0(q, t) + c(q, t)\left[\frac{p}{p-k}(\theta_p - \theta_1) + \frac{k}{k-p}(\theta_k - \theta_i)\right]^2 + d(q, t)d(q, t)\left(\frac{k}{|k-p|}\eta_k - \frac{p}{|k-p|}\eta_p\right)^2 + f(q, t)\left[\frac{p}{p-k}(\theta_p - \theta_1) + \frac{k}{k-p}(\theta_k - \theta_1)\right]\frac{k\eta_k - p\eta_p}{|k-p|}. \quad (9)$$

We substitute the above Taylor series expansions into approximate equation (6), integrate over the angular variables, and equate the coefficients of the powers of the angular variables η_k and $(\theta_k - \theta_1)$ on both sides of

the resulting equations to obtain a set of integrodifferential equations for the unknown functions $v_0(k, t)$, $c(k, t)$, $d(k, t)$, and $f(k, t)$ (it follows from the equation for f that $f \equiv 0$):

$$\left(\partial_t + \nu k^2 - \frac{S}{2}\right)v_0 = ik\left(\frac{L}{2\pi}\right)^3 \times \int_{-\infty}^{\infty} dp\left[\frac{\pi\alpha_1}{2} v_0(p) v_0(k-p)\frac{p^4}{(k-p)^2} + \alpha_1\beta_1 c(p) v_0(k-p)\frac{p^4}{(k-p)^2} + \frac{\pi}{2}\alpha_2 d(p) v_0(k-p)\frac{p^4}{(k-p)^2} + \alpha_1\beta_1 v_0(p)c(k-p)\frac{p^6}{(k-p)^4} + \alpha_1\beta_2 c(p)c(k-p)\frac{p^6}{(k-p)^4} + \alpha_2\beta_1 d(p)c(k-p)\frac{p^6}{(k-p)^4} + \frac{\pi}{2}\alpha_2 v_0(p)d(k-p)\frac{p^6}{(k-p)^4} + \beta_1\alpha_2 c(p)d(k-p)\frac{p^6}{(k-p)^4} + \frac{\pi}{2}\alpha_3 d(p)d(k-p)\frac{p^6}{(k-p)^4}\right], \quad (10)$$

$$\left(\partial_t + \nu k^2 - \frac{S}{2}\right)C + S v_0 = ik^3\left(\frac{L}{2\pi}\right)^3 \times \int_{-\infty}^{\infty} dp \frac{p^6}{(k-p)^6}\left[\alpha_1\frac{\pi}{2} v_0(p)c(k-p) + \alpha_1\beta_1 c(p)c(k-p) + \frac{\pi}{2}\alpha_2 d(k)c(k-p)\right], \quad (11)$$

$$\left(\partial_t + \nu k^2 - \frac{S}{2}\right)d + \frac{S}{2} v_0 = ik^3\left(\frac{L}{2\pi}\right)^3 \times \int_{-\infty}^{\infty} dp \frac{p^6}{(k-p)^6}\left[\frac{\pi}{2}\alpha_1 c(p)d(k-p) + \alpha_1\beta_1 c(p)d(k-p) + \frac{\pi}{2}\alpha_2 d(k)d(k-p)\right]. \quad (12)$$

Here,

$$\alpha_i = \int_{-\pi/2}^{\pi/2} \eta^{2i}\cos\eta\, d\eta, \quad \beta_i = \int_0^{\pi/2}\left(\theta - \frac{\pi}{4}\right)^{2i} d\theta,$$

or, equivalently, $\alpha_1 \approx 0.93$, $\alpha_2 \approx 0.96$, $\alpha_3 \approx 1.29$, $\beta_1 \approx 0.32$, and $\beta_3 \approx 0.12$. If $m = (2\pi/L)\sqrt{3}$ is the absolute value of the smallest wavenumber, then it is clear from physical considerations that the range of wavenumbers determined by the inequalities $|k - p| < m$ and $|p| < m$ should be excluded from the interval of integration because it corresponds to large-scale motions. The amount of computer time required to solve this set of equations is far less than that for the basic equation, because the problem in wavenumber space is one-dimensional even when the grid Reynolds number $Re_L = SL^2/\nu$ is very large. The above set of integrodifferential equations can be nondimensionalized by the transformations $K = kL$, $T = tS/L$, $D = d/(SL)$, $V(K) = v_0/(SL)$, and $C(K) = c/(SL)$. If we also introduce the new variables $G = D/k^4$, $Q = C/k^4$, and $F = V/k^2$, then we can eliminate the wavenumbers from the denominators in the integrands:

$$\left(\partial_T + \frac{1}{Re_L}K^2 - \frac{1}{2}\right)K^2 F = iK(2\pi)^{-3} \times \int_{-\infty}^{\infty} dP\left[\frac{\pi}{2}\alpha_1 F(P)F(K-P)P^6 + \alpha_1\beta_1 Q(p)F(k-p)p^8 + \frac{\pi}{2}\alpha_2 p^8 G(p)F(K-P) + \alpha_1\beta_1 P^8 F(P)Q(K-P) + \alpha_1\beta_2 P^{10}Q(P)Q(K-P)\alpha_2\beta_1 P^{10}G(P)Q(K-P) + \frac{\pi}{2}\alpha_2 P^8 F(P)G(K-P) + \beta_1\alpha_2 P^{10}Q(P)G(K-P) + \frac{\pi}{2}\alpha_3 P^{10}G(P)G(K-P)\right], \quad (13)$$

$$\left(\partial_T + \frac{1}{Re_L}K^2 - \frac{1}{2}\right)K^2 Q + F = iK(2\pi)^{-3} \times \int_{-\infty}^{\infty}\left[\frac{\pi}{2}\alpha_1 P^6 F(P)Q(K-P) + \alpha_1\beta_1 P^8 Q(P)Q(K-P) + \frac{\pi}{2}\alpha_2 P^8 G(P)Q(K-P)\right], \quad (14)$$

$$\left(\partial_T + \frac{1}{Re_L}K^2 - \frac{1}{2}\right)K^2 G + \frac{1}{2}F = iK(2\pi)^{-3} \times \int_{-\infty}^{\infty}\left[\frac{\pi}{2}\alpha_1 P^6 F(P)G(K-P) + \alpha_1\beta_1 P^8 Q(P)G(K-P) + \frac{\pi}{2}\alpha_2 P^8 G(P)G(K-P)\right]. \quad (15)$$

We have thus derived the desired set of three integrodifferential equations for the unknown functions $F(K, T)$, $Q(K, T)$, and $G(K, T)$.

DERIVATION OF THE BASIC EQUATIONS IN PHYSICAL SPACE

We take the Fourier transformation of the functions $F(K, T)$, $G(K, T)$, and $Q(K, T)$ in the variable K according to the formula [10]

$$\Phi(F) \equiv g(\xi) = \int_{-\infty}^{\infty} F(K) e^{-i\xi K} dK. \tag{16}$$

The inverse Fourier transformation is defined as

$$F(K) = (2\pi)^{-1} \int_{-\infty}^{\infty} g(\xi) e^{i\xi K} d\xi. \tag{17}$$

The convolution can be defined as

$$F(K) = \int_{-\infty}^{\infty} F_1(K-P) F_2(P) dP. \tag{18}$$

According to the familiar convolution theorem, we obtain the relationship [10]

$$\Phi(F) = \Phi(F_1)\Phi(F_2). \tag{19}$$

It is known [10] that multiplying the original function by $-iK$ reduces to differentiation of its Fourier transform:

$$\partial_\xi \Phi(F) = \Phi(-iKF(K)). \tag{20}$$

Applying Fourier transformation (15) to both sides of each of Eqs. (12)–(14) and using the above two properties of the Fourier transform, we arrive at the following set of partial differential equations:

$$\left(\partial_T - \frac{1}{Re_L}\partial_\xi^2 - \frac{1}{2}\right)\partial_\xi^2 X = (2\pi)^{-3}\partial_\xi \times \left(-\frac{\pi}{2}\alpha_1 X \partial_\xi^6 X + \alpha_1\beta_1 X \partial_\xi^8 Y + \frac{\pi}{2}\alpha_2 X \partial_\xi^8 Z + \alpha_1\beta_1 Y \partial_\xi^8 X - \alpha_1\beta_2 Y \partial_\xi^{10} Y - \alpha_2\beta_1 Y \partial_\xi^{10} Z + \frac{\pi}{2}\alpha_2 Z \partial_\xi^8 X - \beta_1\alpha_2 Z \partial_\xi^{10} Y - \frac{\pi}{2}\alpha_3 Z \partial_\xi^{10} Z\right), \tag{21}$$

$$\left(\partial_T - \frac{1}{Re_L}\partial_\xi^2 - \frac{1}{2}\right)\partial_\xi^2 Y - X = (2\pi)^{-3}\partial_\xi \times \left(-\frac{\pi}{2}\alpha_1 Y \partial_\xi^6 X + \alpha_1\beta_1 Y \partial_\xi^8 Y + \frac{\pi}{2}\alpha_2 Y \partial_\xi^8 Z\right), \tag{22}$$

$$\left(\partial_T - \frac{1}{Re_L}\partial_\xi^2 - \frac{1}{2}\right)\partial_\xi^2 Z - \frac{1}{2}X = (2\pi)^{-3}\partial_\xi \times \left(-\frac{\pi}{2}\alpha_1 Z \partial_\xi^6 X + \alpha_1\beta_1 Z \partial_\xi^8 Y + \frac{\pi}{2}\alpha_2 Z \partial_\xi^8 Z\right), \tag{23}$$

where $\Phi(F) = X(\xi, t)$, $\Phi(Q) = Y(\xi, t)$, and $\Phi(G) = Z(\xi, t)$ are unknown functions.

On the one hand, the absolute value of the dimensionless wavenumber of turbulent pulsations satisfies the inequality $|k| \geq 2\pi\sqrt{3}$; on the other hand, we have $|\mathbf{K}| = (2\pi/l)\sqrt{n_x^2 + n_y^2 + n_z^2}$, where $n_x = 1, 2, 3, \ldots$, $n_y = 1, 2, 3, \ldots$, and $n_z = 1, 2, 3, \ldots$. Consequently, we get $|\mathbf{K}| = (2\pi/l)\sqrt{3}$, which yields $l = 1$; i.e., in Eqs. (20)–(22), we have $\xi \in [0, 1]$.

Equations (20)–(22) should probably be supplemented with the boundary conditions

$$X(0, T) = Y(0, T) = Z(0, T) = X(1, T) = Y(1, T) = Z(1, T) = 0, \tag{24}$$

$$\partial_\xi X(0, T) = \partial_\xi Y(0, T) = \partial_\xi Z(0, T) = \partial_\xi X(1, T) = \partial_\xi Y(1, T) = \partial_\xi Z(1, T) = 0. \tag{25}$$

Since, under boundary conditions (24) and (25), Eqs. (20)–(22) are degenerate at the boundary, there is no need to impose additional boundary conditions on the higher order derivatives.

The problem of formulating periodic boundary conditions requires separate consideration because Eqs. (21)–(23) have been derived by using Fourier series expansions.

CONCLUSIONS

In the present work, it has been shown that three-dimensional equations for the pulsating components of the velocity can be reduced to a set of one-dimensional approximate equations. As is the case with the familiar original Burgers equation (with a source) [11], the equations obtained can be expected to possess several attractors. In order for the model to be noncontradictory, among these attractors, it is necessary to choose those that maximize the velocity $v^1(\mathbf{k}, t)$ of unstable modes in angular variables. Then, it can be hoped that the spectra obtained of the energy and Reynolds stresses will agree with the experimental spectra and that the approach proposed here will be helpful for subgrid simulation of well-developed turbulence. The author also hopes that the solutions derived will confirm or reject the hypothesis of the "negative diffusion" of the specific rate of dissipation of the turbulent energy [12]. In the limit $Re_L \longrightarrow \infty$, which corresponds to well-developed turbulence, the above model equations and their asymptotic behavior can be investigated numerically. It is possible that the renormalization group methods [13], which are effective in solving problems concerning isotropic turbulence, will also be useful in constructing approximate solutions to the model equations for anisotropic turbulence. Taking into account the second component of the small-scale velocity, v^2, will not considerably complicate the model equations and, presumably, will be needed to describe the cores of coherent vortices—the regions in which the large-scale vorticity predominates over the shear [14].

The approach proposed here can naturally be generalized to magnetohydrodynamic flows and, possibly, to other types of turbulent flows, including those in compressible fluids and gases.

REFERENCES

1. A. S. Monin and A. M. Yaglom, *Statistical Fluid Mechanics*, Vol. 2: *Mechanics of Turbulence* (Gidrometeoizdat, St. Petersburg, 1996; MIT Press, Cambridge, 1975).
2. U. Frisch, *Turbulence. The Legacy of A.N. Kolmogorov* (Cambridge Univ. Press, Cambridge, 1995).
3. G. E. Skvortsov, Vest. Leningr. Gos. Univ. **13** (3), 94 (1979).
4. S. Nazarenko, N. K.-R. Kevlahan, and B. Dubrulle, Physica D **139**, 158 (2000).
5. A. M. Balonishnikov, Zh. Tekh. Fiz. **73** (10), 36 (2003) [Tech. Phys. **48**, 270 (2003)].
6. J. Lee, J. Math. Phys. **16**, 1359 (1975).
7. A. D. Bryuno, *Local Method of Nonlinear Analysis of Differential Equations* (Nauka, Moscow, 1979; Springer-Verlag, Berlin, 1989).
8. H. Haken, *Synergetics: An Introduction* (Springer-Verlag, Berlin, 1978; Mir, Moscow, 1980).
9. V. V. Ivanov, *Method of Computer Calculations: A Handbook* (Naukova Dumka, Kiev, 1986).
10. J. Mathews and R. L. Walker, *Mathematical Methods of Physics* (Benjamin, New York, 1964; Atomizdat, Moscow, 1972).
11. J. M. Bargers, Adv. Appl. Mech. **1**, 171 (1948).
12. A. M. Balonishnikov, Phys. Rev. E **61**, 1390 (2000).
13. L. Ts. Adzhemyan, N. V. Antonov, and A. N. Vasiliev, *Field Theoretical Renolmalization Group in Fully Developed Turbulence* (Gordon and Breach, London, 1999).
14. Y. Dubief and F. Delcayre, J. Turbulence **1**, 2 (2000); http://jot.iop.org.

Translated by O. Khadin

Technical Physics, Vol. 48, No. 11, 2003, pp. 1413–1418. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 53–58.

GAS DISCHARGES, PLASMA

Modeling of a Discharge in Crossed Fields with Account of Collisions and Ionization

A. Yu. Kovalenko and Yu. A. Kovalenko

All-Russia Electrotechnical Institute, Krasnokazarmennaya ul. 12, Moscow, 111250 Russia

e-mail: arkov@mail.ru

Received April 3, 2003

Abstract—A one-dimensional mathematical model is proposed for calculating the characteristics of a discharge in crossed electric and magnetic fields. In particular, the model can be applied for calculating electric propulsion engines with closed (azimuthal) drift of electrons. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

Electric propulsion engines (EPEs) with closed (azimuthal) drift of electrons are based on the concept of acceleration of ions by a self-consistent electric field generated in a plasma because of a sharp decrease in the transverse electron mobility in a magnetic field. A characteristic feature of electric discharge in EPEs is that the discharge is excited in crossed electric and magnetic fields ($E \perp H$) and the ion component of the plasma is not magnetized. The longitudinal component is dominant in the magnetic field, whereas the transverse component is dominant in the electric field (Fig. 1). The Hall current in this case is closed in the azimuthal direction. The presence of this current can cause the external magnetic field to be completely pushed out of the plasma. It is easy to verify that the electron dynamics in the boundary layer between the magnetic field and the cathode plasma is collisionless because, for a characteristic density of neutral atoms on the order of 10^{13} cm^{-3} and a total collision cross section of 3×10^{-15} cm^2, the electron mean free path is about 30 cm, whereas the length of the EPE channel is usually no longer than 5 cm (the electron Larmor radius is ~0.34 cm for a discharge voltage of 100 V and magnetic field of 100 Oe).

FORMULATION OF THE PROBLEM

We consider a quasi-collisional model of a space-charge layer in crossed electric and magnetic fields. A similar problem was first investigated in [1]. In that paper, the case was considered in which all of the plasma particles entering the boundary layer had the same initial velocity directed perpendicular to the layer. However, in contrast to anode-layer engines (ALEs), the ions moved together with the electrons from the cathode plasma, rather than from the anode. The distributions of the potential, magnetic field, and charged particle density and velocity across the layer were found. In particular, it follows from the solutions obtained that the particle density becomes infinite at the anode boundary.

A distinctive feature of ALEs is that the current density of the accelerated ion beam is not limited by the space charge, as was demonstrated in [2]. In [3], the problem was considered of the formation of a space-charge layer in a transverse magnetic field with ions injected from the anode and electrons injected from the cathode. In this case, the current is limited by the space charge and two types of solution are possible that differ in the spatial profiles of the electron velocity and the parameter values of the cathode plasma, from which the electrons arrive at the layer. The first type of solution (*E*-layer) is characterized by the longitudinal acceleration of electrons near the cathode and their deceleration near the anode. A characteristic feature of the second type of solution (*H*-layer) is that the electrons are decelerated in the longitudinal direction throughout the entire layer (from the cathode to the anode). The cathode plasma density in the *E*-layer is higher than that in the *H*-layer by a factor of $\sqrt{\varphi_0/\varphi_n}$, where φ_0 is the voltage drop across the layer and φ_n is the initial energy with which the electrons arrive at the layer from the cathode plasma.

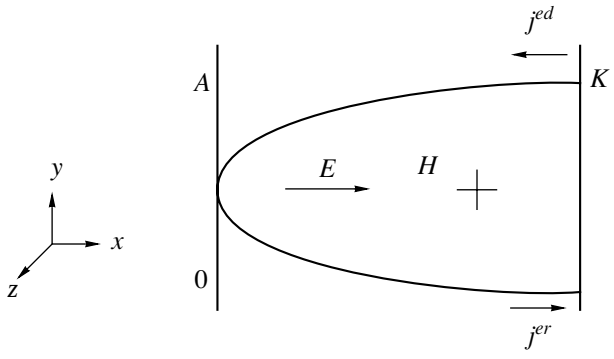


Fig. 1. Space-charge layer.

1063-7842/03/4811-1413$24.00 © 2003 MAIK "Nauka/Interperiodica"

In addition, when the ion and electron current densities far exceed the Langmuir current density, the *E*-layer turns out to be much shorter than the electron Larmor radius. The *H*-layer is close in size to the electron Larmor radius. We note that in [3] the influence of collisions and ionization on the electron and ion motion was ignored. This study is a continuation and further development of [3].

MATHEMATICAL MODEL

Let us consider a space-charge layer formed between a metal anode at a potential φ_0 and a plasma at a zero potential (Fig. 1). An electron flow with an initial electron energy φ_n and current density j_0 arrives at the layer from the cathode plasma.

A metal (equipotential) anode is located immediately near the cutoff point, at which the longitudinal electron velocity is zero ($V_x^e = 0$) and the transverse electron velocity is

$$V_y^e = c^2\left(1 - \frac{1}{\gamma_a^2}\right), \quad \gamma_a = 1 + \frac{e(\varphi_0 + \varphi_n)}{mc^2}, \qquad (1)$$

where c is the speed of light, e is the electron charge, and m is the electron mass.

The electrons that are "cut off" under the action of the transverse magnetic field and have not undergone collisions return to the cathode, thus producing the current density j_c. The electrons undergoing collisions are assumed to be ejected onto the anode [3] and are not taken into account.

A gas flow with the current density $n_g^0 V_g$ (where n_g^0 and V_g are the density and velocity of the neutral particles, respectively) enters the layer through the anode. The gas flow is ionized due to collisions with electrons. As a result, an ion flow and a gas flow with the current density $n_g^c V_g$ enter the cathode plasma (here, n_g^c is the neutral density at the cathode plasma boundary). The magnetic field at the anode is equal to the external field H_0 and decreases toward the cathode because of the presence of a transverse (Hall) electron current. The electric field at the cathode is zero, because the cathode is the boundary of the quasineutral plasma.

Unlike [3], we will consider the continuity equations for the charged particle flows and the gas flow. Moreover, the integration of the equations derived will be performed starting from the anode, which will require the equations describing the problem to be somewhat modified.

The electron flow arising at the cathode reaches the anode and then moves in the opposite direction. Therefore, we can consider two electron flows: the direct flow with the current density j^{ed} and the reverse flow with the current density j^{er}. We assume that the x axis is directed along the reverse electron flow; accordingly, the direct flow is antiparallel to the x axis.

In the one-dimensional problem under study, all variables depend on x only; hence, the continuity equations for j^{ed} and j^{er} take the form

$$\frac{d}{dx}j^{ed} = z_t n^{ed}(x) + z_k n^{ed}(x),$$

$$z_t = \langle \sigma_t v^e \rangle = \sigma_t(V^e(x))V^e(x),$$
$$\frac{d}{dx}j^{er} = -z_t n^{er}(x) - z_k n^{er}(x), \qquad (2)$$

$$z_k = \langle \sigma_k v^e \rangle = \sigma_k(V^e(x))V^e(x),$$

where z_t is the electron–neutral collision frequency, v^e is the electron velocity, z_k is the Coulomb collision frequency, σ_t is the cross section for electron–neutral collisions, σ_k is the cross section for Coulomb collisions of electrons, and $V^e(x)$ is the mean electron velocity at the point x.

By virtue of the energy conservation law and the reversibility of the electron motion in electric and magnetic fields, the electron velocities in the direct and reverse flows at a given point x are the same in magnitude, the velocity components V_y^e coincide in magnitude and direction, and the velocity components V_x^e differ only in direction. With this fact taken into account, it follows from Eqs. (2) that

$$j_x^{ed} j_x^{er} = \text{const} = j_x^e(0)^2,$$

because the densities of the direct and reverse electron flows coincide at $x = 0$.

When solving the problem, we will specify the values of the functions at the anode. However, the value of the total electron current density at the anode $j_x^e(0)$ is not known in advance. It is only known that the electrons arrive at the layer from the cathode plasma with the flow density j_0. By solving the set of Eqs. (2), we can relate the quantities j_0 and $j_x^e(0)$ through a solution for a given value of $j_x^e(0)$, which corresponds to a certain value of the reverse electron current density at the cathode, $j_x^{er}(x_c)$.

It follows from the above considerations that the total electron density at the point x is equal to

$$n_e(x) = n^{ed}(x) + n^{er}(x)$$
$$= j_x^2(0)\left(i^{er}(x) + \frac{1}{i^{er}(x)}\right)/\left|V_x^e(x)\right|, \qquad (3)$$

where

$$i^{er}(x) = \frac{j_x^{er}}{j_x^e(0)}.$$

Thus, the continuity equation for the reverse electron flow is sufficient to close the set of equations.

In [3], ions were assumed to arrive at the layer from the anode surface. Here, we will assume that the gas flow arrives at the layer from the anode surface and is then ionized by electrons.

Due to ionization, the gas flow density decreases as the gas moves from the anode to the cathode. This decrease can be found by integrating the continuity equation

$$\frac{dj_g(x)}{dx} = -z_i(x)n_e(x), \quad j_g(0) = j_g^0, \tag{4}$$

where $j_g(x)$ is the gas flow density, $n_e(x)$ is the electron density, and j_g^0 is the gas flow density at the anode.

The ions produced by the gas ionization in the layer move toward the cathode under the action of the electric field and do not undergo collisions. We also assume that the magnetic field does not affect the ion motion. In this case, the ion space-charge density at a given point is expressed by the integral from the point at which the ions are generated to the observation point.

To simplify the problem and reduce it to solving the set of ordinary differential equations (instead of solving a set of integrodifferential equations), we consider the ion motion using the well-known current tube method. To solve the problem, it is also necessary to use the equations of motion for electrons, Poisson's equation, and equation for the magnetic field. We also make the substitution of variables in the form $V_x^e = dx/dt$. The problem under consideration is time-independent; hence, when writing $V_x^e(x) = dx/dt$, it should be remembered that t is not the time in its usual sense. The ions pass the distance dx over a much longer time than electrons do. So, if dx in the equations for ions and electrons is the same, then we have

$$dx = V_x^+(x)dt^+ = V_x^e dt \quad \text{and} \quad dt^+ = \frac{V_x^e(x)}{V_x^+(x)}dt, \tag{5}$$

where $V_x^+(x)$ is the ion velocity and dt^+ is the "ion" time.

This relation must be used when solving the equation of motion for ions. We used the following normalization:

$$L_0 = \sqrt{\gamma_0^2 - 1}\frac{mc^2}{eH_0}; \quad t_0 = \frac{mc}{eH_0}; \quad \eta = \frac{\varphi}{\varphi_0};$$
$$h = 1 - \frac{H}{H_0}; \quad \gamma_0 = 1 + \frac{e(\varphi_0 + \varphi_n)}{mc^2}. \tag{6}$$

Under the above assumptions, the problem under study can be described by the following set of first-order differential equations:

$$\frac{dv_x^e}{d\tau} = \frac{e_x}{\gamma_0 + 1} + v_y^e(1-h),$$

$$\frac{dv_y^e}{d\tau} = -v_x^e(1-h),$$

$$\frac{dh}{d\tau} = 0.5i_0\left(i + \frac{1}{i}\right)v_y^e,$$

$$\frac{de_x}{d\tau} = \frac{0.5}{\gamma_0 - 1}\left(i_0\left(i + \frac{1}{i}\right) - n^+ v_x^e\right),$$

$$\frac{di}{d\tau} = -\chi_t\left(j_g\left(F_t(\gamma) + \frac{\sigma_i^{\max}}{\sigma_t^{\max}}F(\gamma)\right)\right.$$
$$\left. + 4.0U_g\frac{10^{-8}n^+}{g\sigma_t^{\max}(\gamma-1)^2}\right)iv^e,$$

$$\frac{dj_g}{d\tau} = -kj_g\left(i + \frac{1}{i}\right)v^e F(\gamma), \tag{7}$$

$$\frac{d\eta}{d\tau} = v_x^e e_x,$$

$$\frac{dj_k^+}{d\tau} = gkj_g\left(i + \frac{1}{i}\right)v^e F(\gamma) \quad (\tau_{k-1} < \tau < \tau_k)$$

$$\begin{cases} j_k^+(\tau) = 0 \quad \text{for} \quad \tau < \tau_{k-1}, \\ j_k^+(\tau) = \text{const} = j_k^+(\tau_k) \quad \text{for} \quad \tau \geq \tau_k, \end{cases}$$

$$\frac{dv_k^+}{d\tau} = -\frac{2}{(\gamma_0 + 1)}e_x v_x^e \quad (\tau_k < \tau),$$

$$\frac{dx}{d\tau} = v_x^e,$$

where

$$\chi_0 = n_g\sigma_i^{\max}L_0 i_0\sqrt{\frac{M}{m}} = gk; \quad \chi_T = n_g^0\sigma_t^{\max}L_0;$$

$$\sigma_t(\gamma) = \sigma_t^{\max}F_t(\gamma);$$

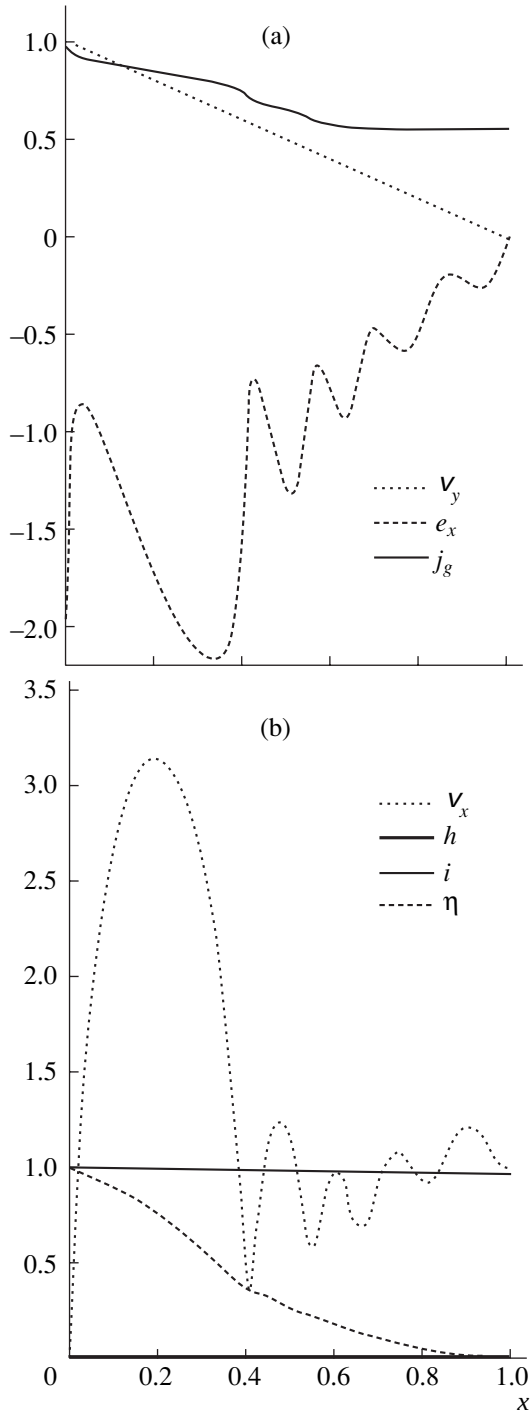


Fig. 2. Simulation results for a Xe flow rate of 0.57 A, U = 150 V, H = 282 Oe, T_e = 1.85 eV (I_p = 0.94 A), and L_0 = 0.148 cm.

$$\gamma_3 = (\gamma_0 - 1)(\gamma_0 + 1);$$

$$\gamma = \sqrt{1 + \gamma_3((v_x^e)^2 + (v_y^e)^2)};$$

$$\sigma_i(\gamma) = \sigma_i^{\max} F(\gamma);$$

$$n^+ = \sum_{i=1}^{n} \frac{j_i^+(x)}{|V_i^+(x)|} j_g(0) = g j_T;$$

$$j^e(0) = i_0 \sqrt{\frac{M}{m}} j_T; \quad j_T = \frac{H^2}{8\pi} \frac{1}{Mc\sqrt{\frac{M}{m}}\sqrt{\gamma_0^2 - 1}}.$$

Here, v_x^e and v_y^e are the x and y components of the electron velocity, respectively; v^e is the total electron velocity; H and h are the magnetic field and the demagnetization, respectively; e_x is the x component of the electric field; i is the electron current; i_0 is the reverse electron current specified at the anode; j_g is the gas flow density; η is the potential; $j_k^+(x)$ is the ion current density in the kth current tube; $V_k^+(x)$ is the ion velocity in the kth current tube at the observation point; and $n^+(x)$ is the ion density.

The transport cross section was specified by a spline, the Coulomb cross section was specified as a function inversely proportional to the square of energy, and the ionization cross section was specified through the Lotz approximation [4]:

$$\sigma_t^{\max} = 5 \times 10^{-15}\ \text{cm}^2 \quad \text{for} \quad E = 2.66\ \text{eV},$$

$$\sigma_t = 8.9 \times 10^{-16}\ \text{cm}^2 \text{ for } E > 10\ \text{eV}, \sigma_k = \frac{10^{-12}}{E^2}\ \text{cm}^2,$$

$$\sigma_i = \sum_{\alpha} a_\alpha q_\alpha \frac{\ln\left(\frac{E}{U_i}\right)}{EU_i}\left[1 - b_\alpha \exp\left(c_\alpha - \left(\frac{E}{U_i} - 1\right)\right)\right], \quad (8)$$

$$\sigma_i^{\max} = 7.5 \times 10^{-16}\ \text{cm}^2,$$

where a_α, b_α, and c_α are constants depending on the type of gas; q_α is the number of equivalent electrons at the shell α; and E is the electron energy.

The mathematical model created on the basis of the above set of equations reduces to numerically solving the set of $2n + 8$ differential equations, where n is the number of the current tubes. The boundary-value problem with the boundary conditions imposed at the anode and cathode was solved using the MathCad package.

The basic input data were the magnetic field strength, the gas flow rate, and the applied voltage. It was also necessary to specify the initial electron velocity and the electron density at the cathode boundary of the layer. The electron velocity is chosen from physical considerations, namely, that the electrons arrive from the cathode plasma with the thermal velocity. Calculations show that variations in T_e within the range 1–5 eV are inessential for the chosen regime of the engine. The electron density is specified using relation (2) through the reverse (originating at the anode) electron current.

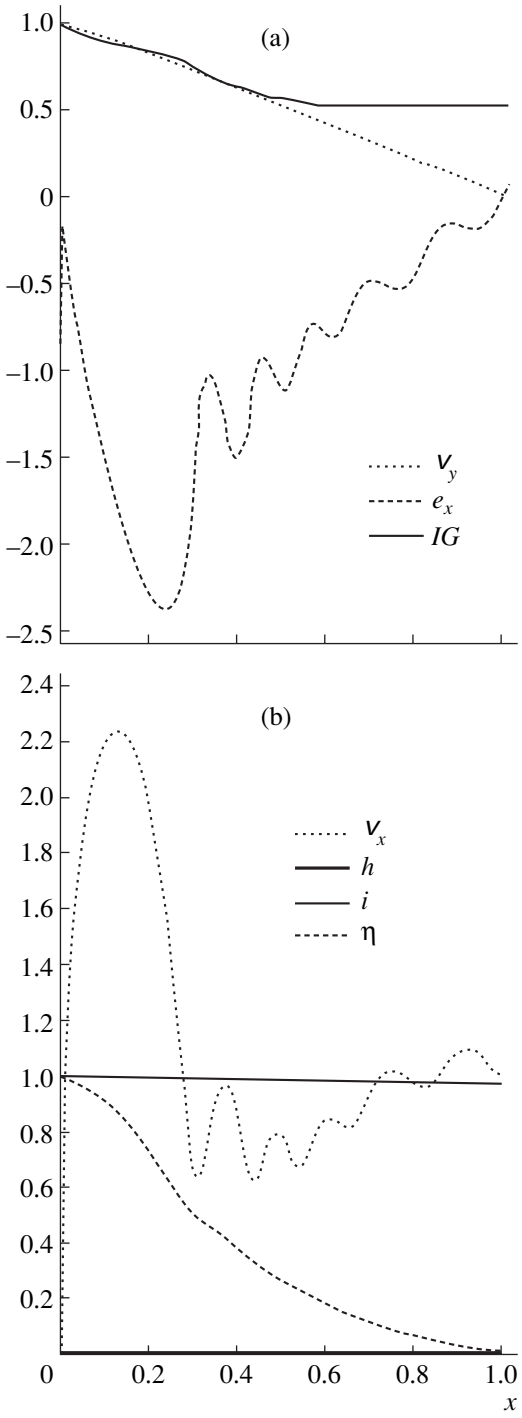


Fig. 3. Simulation results for a Xe flow rate of 0.75 A, U = 95 V, H = 282 Oe, T_e = 2.1 eV (I_p = 1.54 A), and L_0 = 0.12 cm.

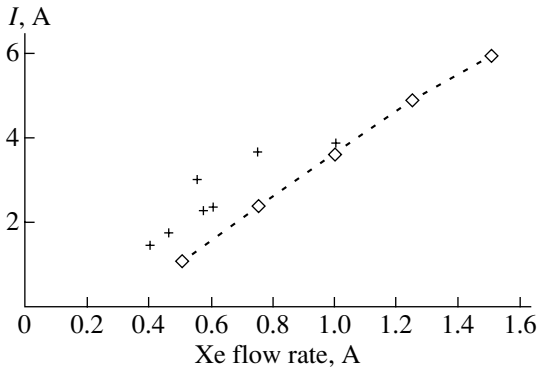


Fig. 4. Discharge current as a function of the Xe flow rate for U = 200 V and H = 282 Oe.

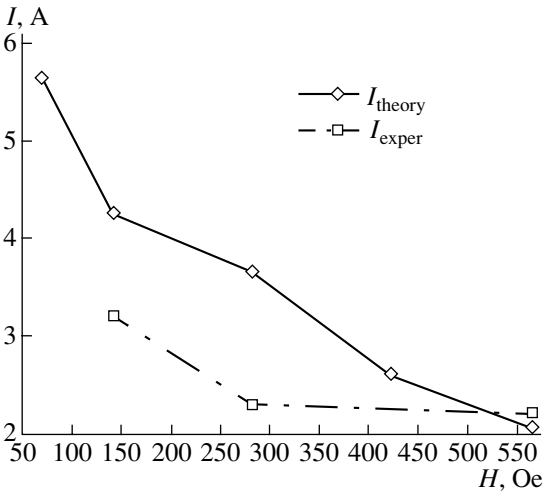


Fig. 5. Discharge current as a function of the magnetic field for U = 200 V, Xe flow rate of 1 A, and T_e = 1.2 eV.

Using the shooting method, we seek a solution that satisfies the initial conditions at the cathode; namely, both the electric field and the transverse electron velocity should vanish at the same point (which determines the boundary of the cathode plasma) at a given initial electron energy and an excess negative charge. The electric field at the anode is a variable parameter too. Its maximum value is easy to estimate from the stringent condition of the magnetron cutoff for electrons at the anode,

$$|e_x| \geq \frac{v_y^e(1-h)(\gamma_0+1)}{c}. \quad (9)$$

The model was tested using the well-known analytical vacuum solution for the case when collisions and ionization in the layer are absent.

ANALYSIS OF CALCULATED RESULTS

The results of calculations for different regimes are shown in Figs. 2 and 3. The relative length of the layer is plotted on the abscissa. For the parameters used in the calculations, this length is close to the electron Larmor radius.

Figure 2 shows the results of calculations for xenon, which correspond to the following experimental parameters: q = 0.57 A, U = 150 V, H = 282 Oe, I = 2.2 A, and $p_k = 4.8 \times 10^{-4}$ torr. It can be seen from the

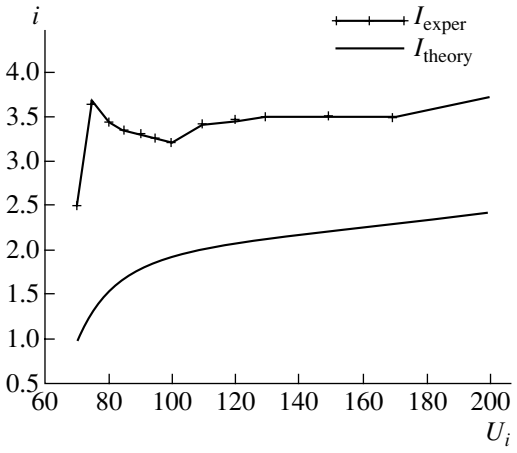


Fig. 6. Calculated and measured current–voltage characteristics for a Xe flow rate of 0.75 A and H = 282 Oe.

figure (curve j_g) that the decrease in the gas flow density is only ~40% and the demagnetization does not exceed several percent. In this case, the calculated density of the cathode plasma is 1.4×10^{11} cm^{-2}.

Figure 3 presents the results of calculations corresponding to the layer parameters used in the analytical collisionless solution [1] and in [3] for the *H*-layer regime, in which the longitudinal electron velocity decreases from the cathode to the anode. In this case, the degree to which the gas is "burnt out" attains 100%, the entire ionization region is located near the anode, and the demagnetization at the cathode boundary reaches 60%.

The difference between the solutions presented in Figs. 2 and 3 consists in different lengths of the region of efficient gas ionization (curves j_g), the degree to which the gas is burnt out, the demagnetization h in the layer, and the layer length. All these results indicate that the solution is of the *H*-layer type.

The calculated and measured dependences of the discharge current on the gas flow rate are shown in Fig. 4. Figure 5 shows the calculated and measured dependences of the discharge current on the magnetic field, and Fig. 6 shows the calculated and measured current–voltage characteristics of the discharge. The calculated dependences in Figs. 4–6 agree qualitatively with the experimental ones. The difference is explained by the fact that doubly ionized ions and scattered electrons were ignored in the model.

CONCLUSIONS

The elaborated one-dimensional quasi-collisionless mathematical model of a space-charge layer formed between a metal anode and plasma in crossed electric and magnetic fields correctly describes the *H*-layer and allows one to calculate its parameters.

The model predicts that, under certain operating conditions, the electric field can almost (but not completely; otherwise, the model is no longer correct) vanish inside a layer near the anode (x = 0). This testifies to the possibility of the generation of the anode plasma, which can be a criterion that the *H*-layer regime changes to the *E*-layer regime.

For more correct calculations, it is necessary to take into account the influence of scattered electrons.

REFERENCES

1. S. Chapman and V. Ferraro, J. Geophys. Res. **57**, 15 (1952).
2. A. V. Zharinov and Yu. S. Popov, Zh. Tekh. Fiz. **37**, 294 (1967) [Sov. Phys. Tech. Phys. **12**, 208 (1967)].
3. M. A. Vlasov, A. V. Zharinov, and Yu. A. Kovalenko, Zh. Tekh. Fiz. **71** (12), 34 (2001) [Tech. Phys. **46**, 1522 (2001)].
4. *Handbook of Physical Quantities*, Ed. by I. S. Grigoriev and E. Z. Meilikhov (Énergoatomizdat, Moscow, 1991; CRC, Boca Raton, 1997).

Translated by N. Larionova

Technical Physics, Vol. 48, No. 11, 2003, pp. 1419–1423. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 59–63.

SOLIDS

Interrelation between Electromagnetic Response Parameters and Impact Excitation Characteristics in Insulators

T. V. Fursa, A. V. Savelev, and K. Yu. Osipov

Tomsk Polytechnical University, pr. Lenina 30, Tomsk, 634050 Russia

e-mail: osipov@tpu.ru

Received March 20, 2003

Abstract—The effect of the impact energy and duration on the parameters of the electromagnetic response from concrete is studied. These two parameters are shown to affect the spectral characteristics of the response. The duration of the first pulse of the response reflects the duration of the impact excitation active stage. An impact excitation energy above 5×10^{-2} J causes irreversible residual strains in concrete. © *2003 MAIK "Nauka/Interperiodica".*

It is known [1] that the mechanical excitation of insulators generates an electromagnetic signal. Experimental and theoretical data for sources and mechanisms of mechanoelectrical transformations in concretes are summarized in [2, 3]. The electromagnetic signal generated by the impact excitation of composites has been shown to have two components: material electrification at the point of impact and mechanoelectrical transformation at the matrix–filler interface. Both components are bound to depend on the impact excitation energy and duration. Our aim is to find this dependence.

Experiments were performed as follows. A test object was struck by a steel ball falling from a certain height. A capacitive sensor was used as a detector. The signal from the detector was applied to a special device interfaced with a PC. This device digitized the signal with a certain step of discretization. The materials under investigation were cement stone and heavy concrete.

As is known [3], when a spherical body of any radius strikes a plane, the nonlinearity coefficient n of the force characteristic $\alpha = bP^n$, where α is the size of the indentation and b is the compliance of the material, is close to unity. Therefore, when describing impact excitation, one can use the formula for the linear force characteristic with a high degree of accuracy.

As follows from the impact theory [3], the duration τ_a of the impact active stage in the case of the linear force characteristic is found by the formula

$$\tau_a = \frac{\pi}{2}\frac{\sqrt{2E_0}b}{V_0}, \quad (1)$$

where E_0 is the kinetic energy of colliding bodies, b is the compliance of the material, and V_0 is the impact velocity.

In our case, the kinetic energy of collision is defined by the potential energy of the ball. Hence, by varying the height of the fall, one can vary the impact energy over wide limits with the impact duration remaining the same, since the duration of the impact active stage does not depend on the impact velocity in the case of the linear force characteristic. The duration τ_a of this stage depends on the mass of the ball, since it follows from formula (1) that

$$\tau_a = \frac{\pi}{2}\sqrt{m}b, \quad (2)$$

where m is the mass of the ball.

When two dissimilar materials are brought into contact, the double electrical layer forms at the contact boundary. The efficiency of its formation depends on the amount of positive and negative charges separated at the boundary. The properties of the double electrical layer, in turn, depend on the properties of contacting materials, the condition of their surfaces, and the contact area. When the ball is indented into the surface, the contact surface area is equal to the surface area of the ball segment indented into the specimen: $S = 2\pi Rh$, where R is the radius of the ball. Hence, if the indentation depth varies linearly, so does the contact surface area. However, during indentation, the contact force increases and reaches a maximum at the end of the process. Therefore, the contact area and, accordingly, the charge of the double electrical layer are bound to grow rapidly at the active stage.

Since we record the variable electric current, that is, the rate of change of the charge at the contact,

$$I(t) = \frac{dQ(t)}{dt},$$

the time variation of the charge itself can be found by integrating the electromagnetic response:

$$Q(t) = \int I(t)dt.$$

1063-7842/03/4811-1419$24.00

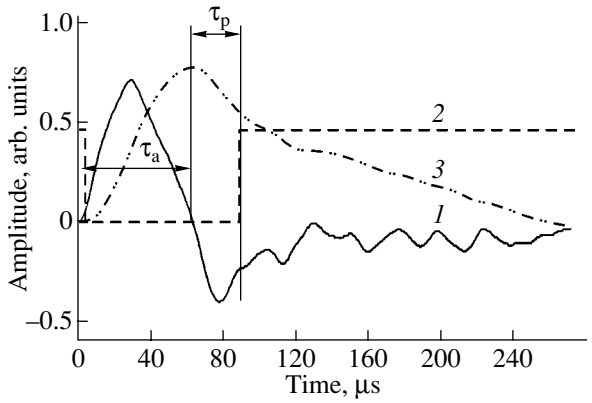


Fig. 1. (*1*) Electromagnetic response to the impact excitation of cement stone, (*3*) its integrated characteristic, and (*2*) variation of the ball potential at the instant of contact with the grounded surface.

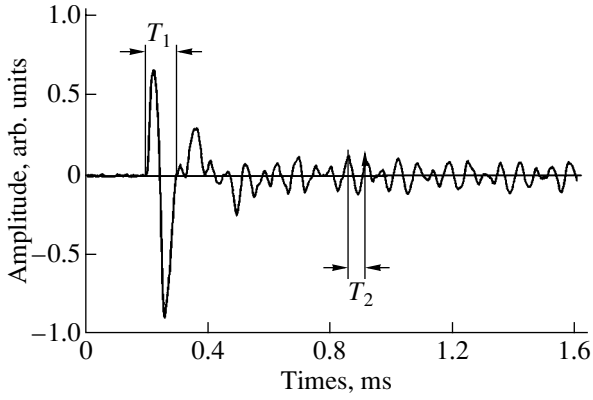


Fig. 2. Typical electromagnetic response to the impact excitation of heavy stone.

Let us analyze the electromagnetic response from a cement stone without a filler in which internal sources of mechanoelectrical transformations (double electrical layers) are weak. To relate the electromagnetic response parameters to the impact excitation duration, we carried out the following experiment. Aluminum foil was laminated on the specimen surface covered by a dielectric lubricant (Litol). The foil was grounded, and the ball and one input (channel) of the measuring device were under a low voltage applied from the power supply. At the instant the ball touches the grounded surface, the voltage dropped to zero and remained zero up to the instant the ball left the surface. In other words, one of the channels measured the total duration of the exciting impact (i.e., the duration of its active and passive stages), while the other measured the electromagnetic response. Figure 1 shows the electromagnetic response (curve *1*) measured synchronously with the total duration of the impact (curve *2*) for the cement stone. The response was numerically integrated to obtain the charge variation at the lubricant–specimen interface (curve *3*). The leading edge time is associated with the separation of positive and negative charges at the contact when the ball is indented into the surface and specifies the active stage duration τ_a. During this (loading) stage, the contact force grows. The time interval τ_a is the duration of the first pulse of the response (curve *1*). Note that the impact in the absence of the foil would generate a solitary response (pulse) of similar shape but slightly different amplitude because of the different electrical properties of concrete and Litol.

The loading stage is followed by the stage of relief (passive stage). In the course of the passive stage (its duration is τ_p), the contact force diminishes and the pulse amplitude drops, since the decrease in the contact surface area changes the width and charge of the double electrical layer. That is, the shape of the pulse obtained by integrating the electromagnetic response reflects the variation of the force characteristic of the impact excitation.

From Fig. 1, it is seen that the rise time of the integrated signal, which is related to the active stage duration, equals 64 μs. Let us estimate the passive stage duration and compare a predicted value with that found from the electromagnetic response parameters. To do this, we will use the basic formulas from the impact theory [3]. From the duration of the first pulse of the electromagnetic response, we find the passive stage duration with the formula $\tau_p = 1.65 P_m^{-1/6}\sqrt{mb}$, where $P_m = V_0\sqrt{m/b}$ is the maximal load and b is the compliance, which is determined from formula (1) for a given τ_a. The calculated and experimental values of τ_p are found to be 27 and 29 μs, respectively, which is one more verification of our reasoning.

Impact excitation not only causes electrification due to the triboelectric effect but also acoustically generates double electrical layers at the boundaries between internal inhomogeneities [5]. Figure 2 shows a typical electromagnetic response from impact-excited heavy concrete with large-area adhesive boundaries between the cement matrix and the binder, where extended double electrical layers may form. Unlike binder-free cement stone (Fig. 1), the electromagnetic response from heavy concrete is of distinct oscillatory character because of the acoustic generation of double electrical layers. In Fig. 2, the period T_1 of the first pulse differs from that of subsequent oscillations (T_2). This is because the first pulse is related to the impact parameters, while the subsequent variation of the electromagnetic response is due to natural acoustic oscillations, which generate double electrical layers at the matrix–binder interface.

To trace the effect of the impact duration on the response parameters, we used steel balls weighing 0.73, 1.45, 3.26, and 7.41 g. To exclude the effect of the exci-

tation energy, the balls fell from different heights so that the excitation energy was the same.

The first pulse duration vs. square root of mass dependence for heavy concrete struck by balls of different mass is demonstrated in Fig. 3. This dependence is seen to be linear in complete accordance with the theory of impact with linear force characteristic:

$$\tau_a = \frac{\pi}{2}\sqrt{mb}.$$

This corroborates the assumption that the duration of the first pulse of the response defines the active stage duration.

It should be noted that wave processes may be disregarded in the case of light balls. For a ball of diameter 7 mm, the time of contact by means of an excitation wave is estimated as 3 μs, while the active stage for the same ball lasts 36 μs.

However, after the ball has bounced back from the surface and even at the passive stage of impact, charge separation may partially persist; then, the trailing edge of the pulse obtained by numerical integration describes the force characteristic at the passive stage inadequately.

We believe that mechanoelectrical transformations in impact-excited composites occur when an acoustic wave changes the width of the matrix–inclusion interface. This, in turn, changes the dipole moment of the double electrical layer at the interface, i.e., generates a displacement current, which is recorded by the measuring system. The displacement is due to impact-induced acoustic vibrations.

Thus, the variation of the excitation duration and, hence, of the excitation spectrum is bound to change the spectral characteristics of the electromagnetic response. The spectral characteristics of the electromagnetic responses from the same specimen of heavy concrete, which was struck by steel balls of various masses, are shown in Fig. 4. The electromagnetic response spectrum is seen to vary with impact duration, which counts in favor of the above-suggested mechanism behind mechanoelectrical transformations in the materials.

The effect of the impact energy on the electromagnetic response parameters was studied by throwing a steel ball of weight 7.41 g from various heights. It was found that the first pulse duration does not depend on the impact energy, while the efficiency of mechanoelectrical transformations is energy dependent. This also supports our basic ideas.

Figure 5 shows the dependence of the first pulse amplitude on the impact energy. As the energy grows, so does the amplitude of the response. The dependence is first linear but deviates from linearity after the energy has reached 3.5×10^{-2} J (curve *1*). The nonlinearity may be attributed to microplastic strains. It turned out that impact energies exceeding 5×10^{-2} J (curve *2*) cause irreversible residual strains in concrete, as demonstrated by hysteresis in the energy dependence of the electromagnetic response amplitude.

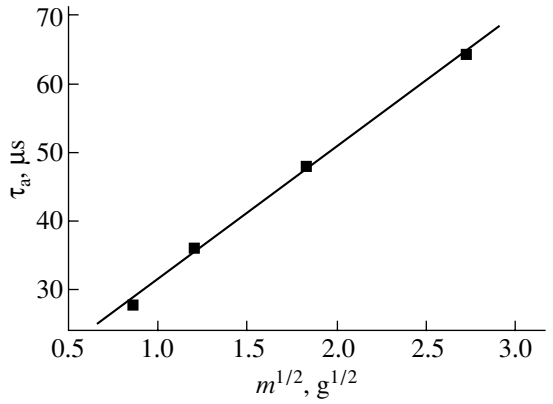


Fig. 3. First pulse duration vs. square root of ball mass for the heavy concrete specimen.

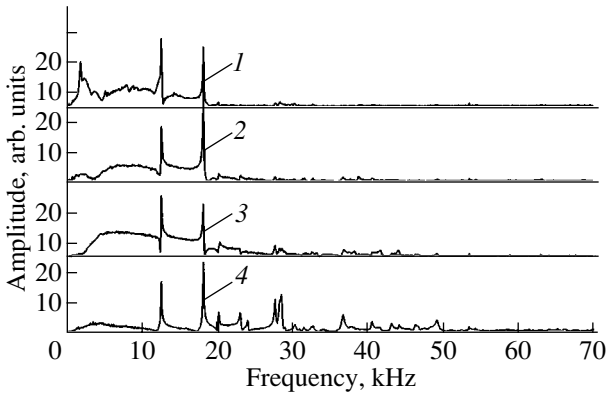


Fig. 4. Spectral characteristics of the electromagnetic response under impact excitation by a ball of mass (*1*) 7.41, (*2*) 3.26, (*3*) 1.75, and (*4*) 0.73 g.

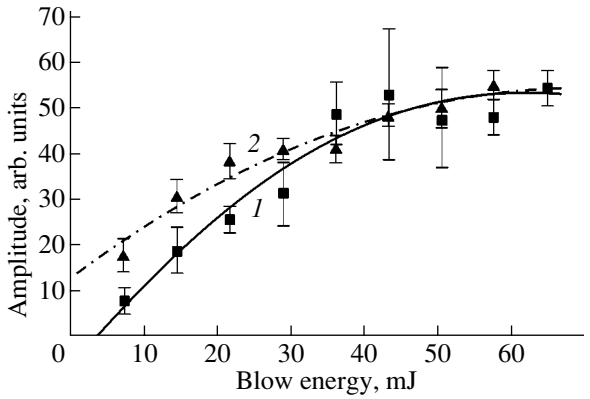


Fig. 5. First (basic) pulse amplitude vs. impact excitation energy.

Figure 6 shows the spectral characteristics of the electromagnetic responses that correspond to the linear

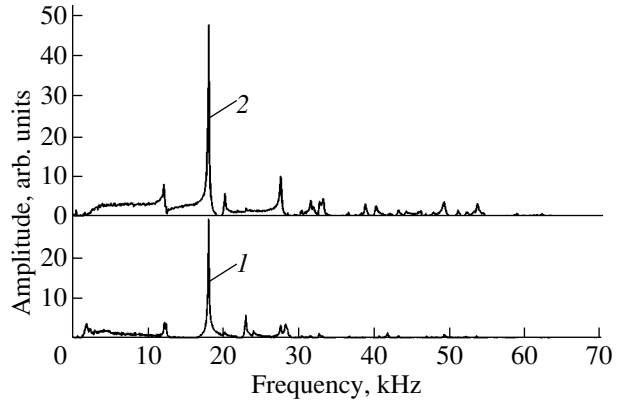


Fig. 6. Amplitude–frequency characteristics of the electromagnetic response that correspond to the (*1*) linear and (*2*) nonlinear strain range for concrete.

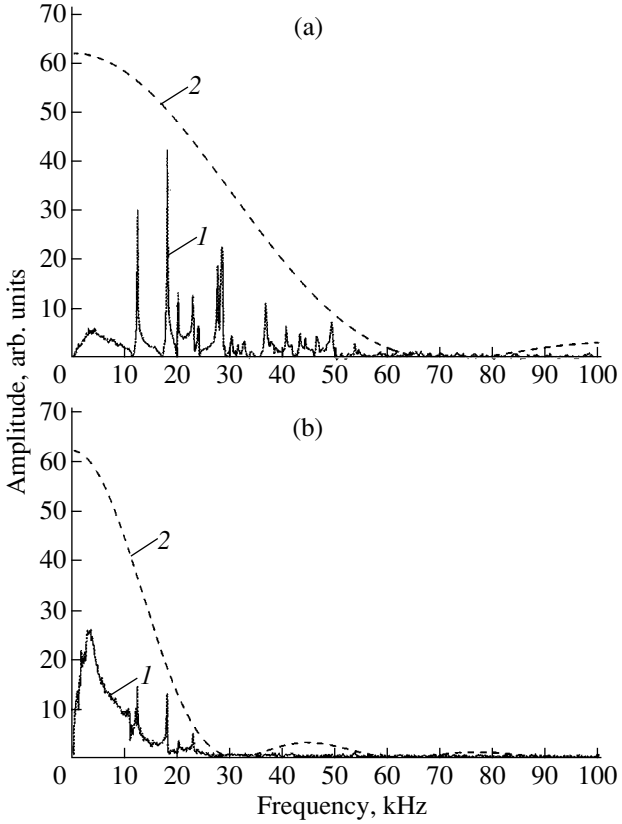


Fig. 7. (*1*) Electromagnetic response spectra and (*2*) excitation spectra calculated from the first pulse duration. The ball mass is (a) 0.73 and (b) 7.41 g.

(curve *1*) and nonlinear (curve *2*) strain portions of the energy dependence, which is shown in Fig. 5. In the latter portion, the response amplitude grows, which is accompanied by the appearance of additional peaks, which may be related to material failure at the point of impact.

Let us trace the effect of the first pulse duration on the spectral characteristics of the response. Note that all tests were made on the same heavy concrete specimen and with the same geometry of the experiment. Under the assumption that the impact duration equals the duration of the leading edge of the integrated electromagnetic pulse, we calculated the excitation spectrum, approximating the exciting pulse by an irregular triangle. The passive stage duration was calculated theoretically with the approach mentioned above, since the superposition of the electromagnetic pulse due to the triboelectric effect and the signal from internal sources makes an estimate of the passive stage duration ambiguous.

Figure 7 shows the electromagnetic response spectra (curve *1*) and the excitation spectra calculated from the first pulse duration (curve *2*) for balls weighing (a) 0.73 and (b) 7.41 g. It is seen that the electromagnetic response spectrum falls into the range of the excitation spectrum calculated from the first pulse duration. As the excitation duration decreases, higher frequency peaks arise in the electromagnetic response spectrum. If the leading edge duration remains the same and the excitation energy grows, the amplitudes of the peaks in the electromagnetic response spectral characteristic increase. Consequently, changes in the impact energy and duration modify the spectral characteristics of the electromagnetic response from concrete.

The results presented in this paper may be used for determining the dynamic properties of the collision process, such as the compliance of the material and the nonlinearity coefficient of the force characteristic in view of dynamic loading. The nonlinearity coefficient n is found by approximating the leading edge of the integrated electromagnetic response by a power function using a routine computer program. The exponent of this power function gives the coefficient n. The coefficient b is given by the formula

$$b = \frac{4\tau^2}{m\pi^2},$$

where τ is the duration of the electromagnetic response first pulse and m is the mass of the ball.

Thus, basically one can find the empiric coefficients characterizing the material's elastoplastic properties from the electromagnetic response to impact excitation.

At present, a method for nondestructive inspection of the material's strength is being developed in the Tomsk Polytechnical University. This method is based on the use of the amplitude–frequency characteristics of the electromagnetic response to impact excitation. It follows from our study that the accuracy of this method can be improved if the energy and spectrum of the exciting pulse is taken into consideration. Further

investigation will be aimed at working out an algorithm for the electromagnetic inspection of material quality in view of the above experimental data.

REFERENCES

1. Yu. P. Malyshkov, V. F. Gordeev, V. P. Dmitriev, *et al.*, Zh. Tekh. Fiz. **54**, 336 (1984) [Sov. Phys. Tech. Phys. **29**, 197 (1984)].
2. V. L. Chakhlov, Yu. P. Malyshkov, V. F. Gordeev, *et al.*, Izv. Vyssh. Uchebn. Zaved., Stroit., Nos. 5–6, 54 (1995).
3. G. S. Batuev, Yu. V. Golubkov, A. K. Efremov, and A. A. Fedosov, *Engineering Methods for Studying Shock Processes* (Mashinostroenie, Moscow, 1977).
4. T. V. Fursa, Zh. Tekh. Fiz. **71** (7), 53 (2001) [Tech. Phys. **46**, 840 (2001)].
5. T. V. Fursa, N. N. Khorsov, and E. A. Baturin, Zh. Tekh. Fiz. **69** (10), 51 (1999) [Tech. Phys. **44**, 1175 (1999)].
6. T. V. Fursa and N. N. Khorsov, Defektoskopiya, No. 2, 68 (2000).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1424–1428. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 64–68.

SOLIDS

Modulation Instability of Electromagnetic Excitations in the Nonlocal Josephson Electrodynamics of Thin Magnetic Superconducting Films

A. I. Lomtev

Galkin Physicotechnical Institute, National Academy of Sciences of Ukraine, Donetsk, 83114 Ukraine
e-mail: lomtev@kinetic.ac.donetsk.ua

Received November 22, 2002; in final form, March 25, 2003

Abstract—The modulation instability of plane finite-amplitude nonlinear waves with the Josephson frequency is studied for a Josephson contact in thin nonmagnetic and magnetic (two-dimensional) superconducting films. Dispersion relations for the increments of small-amplitude perturbations are derived. Modulation instability is shown to develop in the finite range $0 < Q < Q_b$ of the wavevectors of amplitude perturbations for the nonmagnetic films and for any perturbation wavevector $0 < Q < \infty$ for the magnetic films. © *2003 MAIK "Nauka/Interperiodica".*

1. Despite a large number of magnetic superconductors with unique properties [1–3] known today, they are continuing to attract researchers' attention. A combination of magnetism and superconductivity has been found not only in ternary compounds [4] but also in HTSCs, such as REBaCuO, RECuO, etc., where RE is a rare-earth ion. A basic property of HTSCs is the strong antiferromagnetic correlation of copper spins in CuO_2 planes in the superconducting state [5].

Wave instability in various nonlinear systems and media has been studied over a long period of time but continues to be of interest to the present day [6, 7]. It is known that a nonlinear wave can be compressed in both the transverse and longitudinal direction with respect to the direction of wave propagation. Examples are self-focusing of light predicted by Askar'yan [8] and the splitting of a wave into packets with the subsequent self-compression of the packets (modulation instability), which was first investigated by Lighthill [9]. The modulation instability of electromagnetic waves in optical fibers is described by the instability of solutions to the nonlinear Schrödinger equation [10] and, in distributed Josephson junctions, by the instability of solutions to the sine-Gordon equation [11, 12]. The phenomenon of modulation instability is of both theoretical and applied interest. For example, it is used for generating a chain of ultrashort optical pulses with a high repetition rate [10] and developing advanced logic [13].

Upon studying modulation instability, it is often necessary to consider spatial nonlocal modifications of the nonlinear Schrödinger equation [14] and sine-Gordon equation [15–26].

It has been shown [15, 16] that the effects of spatial nonlocality may be essential even in Josephson contacts made in thick superconductors ($d \gg \lambda$, where d is the film thickness and λ is the London penetration depth), i.e., in the situations that were previously analyzed in the local approximation. In the opposite case (very thin films: $d \ll \lambda$), the local limit is absent, while the spatial nonlocality is significant, thereby becoming the decisive factor. Associated equations have been derived and analyzed in [17–20]. As was demonstrated in [21–23], temporal nonlocality is also of significance, along with spatial nonlocality, for the Josephson electrodynamics of thin (both two- and three-dimensional) magnetic superconducting films. Josephson junctions between two superconducting layers that have a finite thickness in the direction orthogonal to the magnetic field of vortices have been studied in [24]. In [25, 26], Josephson junctions made in butted films and beveled junctions that have a finite thickness in the direction of the magnetic field of vortices have been studied for an arbitrary ratio d/λ.

Because of the different geometries considered in [15–26], the equations of Josephson electrodynamics differ in the form of the kernel of an integral operator that describes the effect of spatial nonlocality. However, in all the works cited, the spatial nonlocality of the equations for phase difference arises because of the field nonlocal joining at the interface and in the superconductor. This is a common reason for spatial nonlocality in the electrodynamics of Josephson contacts, where nonlocality becomes a rule rather than an exception.

Modulation instability in the spatially nonlocal Josephson electrodynamics of a contact between bulk superconductors was first considered in [15]. It was shown that the growth of small amplitude and phase perturbations leads to the modulation instability of an electromagnetic wave with a constant (finite) amplitude, causing a nonlinear frequency shift and a linear

1063-7842/03/4811-1424$24.00

mode dispersion. Spatial nonlocality was found to suppress modulation instability. The modulation instability of a plane nonlinear electromagnetic wave with a finite amplitude and Josephson frequency for a Josephson contact between bulk superconductors was studied in [27]. The instability was caused by the growth of small amplitude perturbations and led to the splitting of the wave into wave packets.

2. One nonlinear system where modulation instability may be observed is a Josephson junction in an ultrathin magnetic (two-dimensional) superconducting film ($d \ll \lambda$). In this system, the phase difference $\varphi(x, t)$ across the junction is described by the nonlinear integro-differential sine-Gordon equation with spatial and temporal nonlocality [21, 22]:

$$\sin\varphi(x,t) + \frac{\beta}{\omega_J^2}\frac{\partial\varphi(x,t)}{\partial t} + \frac{1}{\omega_J^2}\frac{\partial^2\varphi(x,t)}{\partial t^2} = I_J\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dt'K\left(\frac{x-x'}{2\lambda_{\text{eff}}}, t-t'\right)\frac{\partial^2\varphi(x',t')}{\partial x'^2}. \quad (1)$$

Here, ω_J is the Josephson frequency, β is the dissipation factor, $I_J = \lambda_J^2/\lambda$, λ_J is the Josephson penetration depth, $\lambda_{\text{eff}} = \lambda^2/d$ is the effective penetration depth, and the nonlocal (in spatial and temporal variables) integral kernel

$$K\left(\frac{x-x'}{2\lambda_{\text{eff}}}, t-t'\right)$$

has the form

$$K\left(\frac{x-x'}{2\lambda_{\text{eff}}}, t-t'\right) = \int_0^{\infty}\frac{dq}{\pi}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{2\lambda_{\text{eff}}J_0[q(x-x')]\exp[-i\omega(t-t')]}{\mu(\mathbf{q},\omega) + 2q\lambda_{\text{eff}}}, \quad (2)$$

where $J_0(qx)$ is the zero-order Bessel function. The temporal nonlocality of Eq. (1) is due to the frequency dispersion of the permeability $\mu(q, \omega)$ and, in essence, is related to delay processes.

Since $\lambda \gg a$, where a is the lattice constant, it is natural to describe the magnetic subsystem in hydrodynamic terms. In the paramagnetic range of temperature, the permeability is given by [28]

$$\mu(\mathbf{q},\omega) = 1 + 4\pi\chi_0\frac{iDq^2}{\omega + iDq^2}, \quad (3)$$

where χ_0 is the static magnetic susceptibility,

$$D = (1/3)(2\pi)^{1/2}Ja^2[s(s+1)]^{1/2}$$

is the coefficient of spin diffusion for two-dimensional Heisenberg magnetics [29], J is the intralayer exchange parameter, and s is the spin.

Consider the evolution of nonlinear waves with the Josephson frequency ω_J and a small but finite amplitude (like a breather) in the junction. We represent the phase difference $\varphi(x, t)$ as

$$\varphi(x,t) = u(x,t)\exp(-i\omega_J t) + u^*(x,t)\exp(i\omega_J t), \quad |u(x,t)| \ll 1. \quad (4)$$

In the nondissipative limit ($\beta = 0$), we take into account only the lowest nonlinearity order at the fundamental frequency ω_J and assume that the amplitude $u(x, t)$ slowly varies with time. Then, the following inequality is valid:

$$\left|\partial^2 u(x,t)/\partial t^2\right| \ll 2\omega_J\left|\partial u(x,t)/\partial t\right|.$$

By substituting field (4) into Eq. (1), we obtain an expression for the amplitude $u(x, t)$:

$$i\frac{2}{\omega_J}\frac{\partial u(x,t)}{\partial t} + \frac{1}{2}|u(x,t)|^2u(x,t) + I_J\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dt'K\left(\frac{x-x'}{2\lambda_{\text{eff}}}, t-t'\right) \times \exp[i\omega_J(t-t')]\frac{\partial^2 u(x',t')}{\partial x'^2} = 0. \quad (5)$$

This equation is a nonlinear Schrödinger equation with spatial and temporal nonlocality. It has an exact solution in the form of a plane nonlinear wave with a constant (in space and time) amplitude A:

$$u_0(t) = A\exp(iA^2\omega_J t/4), \quad A \ll 1. \quad (6)$$

Let us analyze the stability of this solution. The decomposition of plane wave (6) can be judged from the growth of its small perturbations. To do this, we assume that a random small amplitude perturbation

$$u(x,t) = [A + \psi(x,t)]\exp(iA^2\omega_J t/4), \quad |\psi(x,t)| \ll A \quad (7)$$

arises.

From Eq. (5), we obtain a linear equation for the small perturbation $\psi(x, t)$:

$$i\frac{2}{\omega_J}\frac{\partial\psi(x,t)}{\partial t} + \frac{1}{2}A^2[\psi(x,t) + \psi^*(x,t)] + I_J\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dt'K\left(\frac{x-x'}{2\lambda_{\text{eff}}}, t-t'\right) \times \exp[i\omega_J(1-A^2/4)(t-t')]\frac{\partial^2\psi(x',t')}{\partial x'^2} = 0. \quad (8)$$

Assuming that in (8) $\psi(x, t) = v(x, t) + iw(x, t)$, we find a set of equations for the real and imaginary parts

of the perturbation:

$$\frac{2}{\omega_J}\frac{\partial v(x,t)}{\partial t} + I_J\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dt' K\left(\frac{x-x'}{2\lambda_{\text{eff}}}, t-t'\right)$$
$$\times \exp[i\omega_J(1-A^2/4)(t-t')]\frac{\partial^2 w(x',t')}{\partial x'^2} = 0,$$
$$-\frac{2}{\omega_J}\frac{\partial w(x,t)}{\partial t} + A^2 v(x,t) \qquad (9)$$
$$+ I_J\int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dt' K\left(\frac{x-x'}{2\lambda_{\text{eff}}}, t-t'\right)$$
$$\times \exp[i\omega_J(1-A^2/4)(t-t')]\frac{\partial^2 v(x',t')}{\partial x'^2} = 0.$$

For perturbations of the form (arbitrary perturbations can be represented as a superposition of such fields)

$$v(x,t) = V(Q,\Omega)\exp[i(Qx-\Omega t)],$$
$$w(x,t) = W(Q,\Omega)\exp[i(Qx-\Omega t)] \qquad (10)$$

that propagate along a Josephson junction with a wavevector Q and frequency Ω, we obtain from (9) a dispersion relation $\tilde{\Omega} = \tilde{\Omega}(\tilde{Q})$:

$$\tilde{\Omega}^2 = \frac{L}{2\pi}\tilde{Q}^2 J(\tilde{Q},\tilde{\Omega})\left[\frac{2L}{\pi}\tilde{Q}^2 J(\tilde{Q},\tilde{\Omega}) - A^2\right], \qquad (11)$$

where

$$J(\tilde{Q},\tilde{\Omega}) = \int_0^{\infty}dx\left[1+\tilde{Q}\cosh x + \frac{4\pi\chi_0\tilde{Q}^2\cosh^2 x}{-i\eta(\tilde{\Omega}+1-A^2/4)+\tilde{Q}^2\cosh^2 x}\right]^{-1}. \qquad (12)$$

In (11) and (12), $\tilde{Q} = 2Q\lambda_{\text{eff}}$, $\tilde{\Omega} = \Omega/\omega_J$, $L = I_J/2\lambda_{\text{eff}}$, $\eta = \omega_J/\Omega_{\text{eff}}$, and $\Omega_{\text{eff}} = D/(2\lambda_{\text{eff}})^2$.

Dispersion relation (11), which involves $\tilde{\Omega}(\tilde{Q})$ in implicit form, has, in view of (12), a complex solution: $\tilde{\Omega}(\tilde{Q}) = \mathrm{Re}\tilde{\Omega}(\tilde{Q}) + i\mathrm{Im}\tilde{\Omega}(\tilde{Q})$. With $\mathrm{Im}\tilde{\Omega}(\tilde{Q}) > 0$, small perturbations of amplitude (10) will grow with time, causing the modulation instability of plane nonlinear electromagnetic wave (6).

3. In the case of a nonmagnetic superconducting film with $\chi_0 = 0$, integral (12) is the function of only the wavevector $\tilde{Q}$ (see also [20]):

$$J(\tilde{Q}) = \frac{1}{2\sqrt{1-\tilde{Q}^2}}\ln\frac{1+\sqrt{1+\tilde{Q}^2}}{1-\sqrt{1-\tilde{Q}^2}} \quad \text{for} \quad \tilde{Q}\le 1,$$
$$J(\tilde{Q}) = \frac{2}{\sqrt{\tilde{Q}^2-1}}\arctan\frac{\sqrt{\tilde{Q}^2-1}}{1+\tilde{Q}} \quad \text{for} \quad \tilde{Q}\ge 1. \qquad (13)$$

In this case, dispersion relation (11) passes into an expression that involves $\tilde{\Omega}(\tilde{Q})$ in explicit form:

$$\tilde{\Omega}^2(\tilde{Q}) = \frac{L}{2\pi}\tilde{Q}^2 J(\tilde{Q})\left[\frac{2L}{\pi}\tilde{Q}^2 J(\tilde{Q}) - A^2\right]. \qquad (14)$$

If (14) has a positive solution $\mathrm{Im}\tilde{\Omega}(\tilde{Q}) > 0$ in the finite range of wavevectors $0 < \tilde{Q} < \tilde{Q}_b$, the perturbation grows and modulation instability develops in this range. For $\tilde{Q} \ge \tilde{Q}_b$, $\mathrm{Im}\tilde{\Omega}(\tilde{Q}) \equiv 0$ and the wave is stable. The boundary wavevector $\tilde{Q}_b$ is found from the expression

$$\tilde{Q}_b^2 J(\tilde{Q}_b) = \frac{\pi A^2}{2L}. \qquad (15)$$

By the range of modulation instability, we mean the range of the wavevectors of small amplitude modulations where the amplitude increment $\mathrm{Im}\tilde{\Omega}(\tilde{Q})$ is nonzero and positive.

Fig. 1a shows the numerically calculated finite ranges $0 < Q < Q_b$ of modulation instability (in terms of the wavevector Q of small amplitude modulations) for plane nonlinear electromagnetic wave (6), which oscillates with the Josephson frequency ω_J, in the case of a Josephson junction made in a thin nonmagnetic superconducting film for a fixed amplitude A and three values of the parameter L.

As follows from numerical analysis, for a $\chi_0 \propto 10^{-2}$ (such values of the static magnetic susceptibility are typical of ternary compounds and HTSCs containing rare-earth ions near the magnetic ordering temperature $T_N \propto 1$K) and $\eta = 1$, dispersion relation (11) in view of (12) always (i.e., for any value of the wavevector $\tilde{Q}$) yields a positive perturbation increment $\mathrm{Im}\tilde{\Omega}(\tilde{Q})$. Figure 1b shows the wavenumber dependence of the amplitude perturbation increment for plane nonlinear electromagnetic wave (6), which oscillates with the Josephson frequency ω_J, in the case of a thin magnetic film for a fixed amplitude A and three values of the parameter L. In a narrow wavevector range near $\tilde{Q} = \tilde{Q}_b$, the dependence $\mathrm{Im}\tilde{\Omega}(\tilde{Q})$ has a crossover. Therefore, two ranges of modulation instability that turn into each other can be distinguished: the range of strong

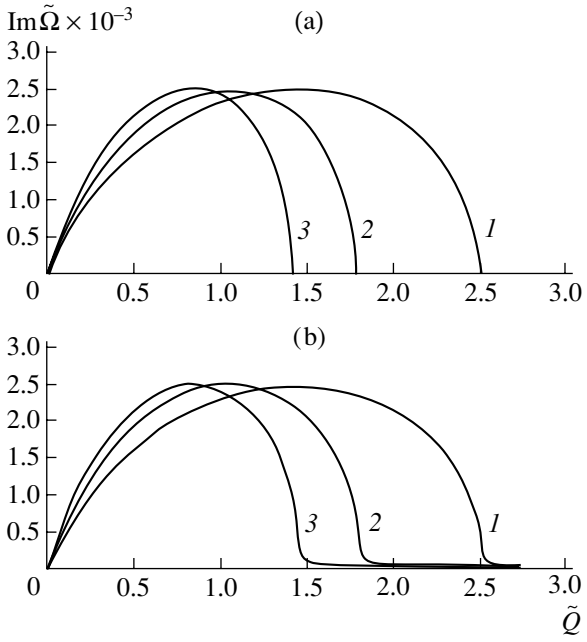


Fig. 1. (a) Reduced amplitude perturbation increment vs. reduced wavevector for plane nonlinear electromagnetic wave (6) in a thin nonmagnetic superconducting film for $A = 0.1$ and $L = (1)\ 0.5 \times 10^{-2}$, $(2)\ 0.75 \times 10^{-2}$, and $(3)\ 1 \times 10^{-2}$. (b) The same for a thin magnetic (two-dimensional) superconducting film.

instability $0 < \tilde{Q} < \tilde{Q}_b$, where $\mathrm{Im}\tilde{\Omega}(\tilde{Q}) \propto 10^{-3}$ for $A = 0.1$, and the range of weak instability $\tilde{Q} > \tilde{Q}_b$, where $\mathrm{Im}\tilde{\Omega}(\tilde{Q}) \leq 10^{-4}$. As the wavevector grows, the perturbation increment tends to zero: $\mathrm{Im}\tilde{\Omega}(\tilde{Q}) \longrightarrow 0$ with $\tilde{Q} \longrightarrow \infty$.

The maximal value of the perturbation increment for both magnetic and nonmagnetic films,

$$(\mathrm{Im}\tilde{\Omega}(\tilde{Q}_m))_{\max} = A^2/4, \qquad (16)$$

is achieved when $\tilde{Q} = \tilde{Q}_m$, where $\tilde{Q}_m$ is a root of the equation

$$\tilde{Q}^2 J(\tilde{Q}) = \frac{\pi A^2}{4L}. \qquad (17)$$

As modulation instability develops, plane nonlinear wave (6) with the Josephson frequency turns into a chain of pulses, small-amplitude breathers, whose repetition rate depends on the modulation period $L_0 = 2\pi/Q$ of the initial wave.

Thus, in the nondissipative limit for a Josephson junction in a thin nonmagnetic superconducting film, modulation instability caused by the growth of small amplitude modulations develops in a finite range of wavevectors $0 < Q < Q_b$. In the same limit for a magnetic superconducting film, temporal nonlocality due to the frequency dispersion of the permeability and spin wave diffusion in the magnetic subsystem causes the modulation instability of plane nonlinear electromagnetic waves in the range $Q > Q_b$, where they are stable when propagating in the nonmagnetic film.

It should be noted that the frequency $\omega(k)$ in the theory of the Josephson junction must be lower than the limiting frequency, which depends on the energy gap width $\Delta(T)$.

Experimentally, modulation instability can be observed in long Josephson junctions consisting of thin nonmagnetic or magnetic superconducting films upon the excitation of waves with a small but finite amplitude that oscillate with the Josephson frequency.

ACKNOWLEDGMENTS

The author thanks Yu.V. Medvedev and I.B. Krasnyuk for valuable discussion and encouragement, A.N. Artemov and S.M. Orel for assistance in the numerical calculation, and A.É. Filippov for graphical assistance.

REFERENCES

1. A. I. Buzdin, L. N. Bulaevskiĭ, M. L. Kulich, and S. V. Panyukov, Usp. Fiz. Nauk **144**, 597 (1984) [Sov. Phys. Usp. **27**, 927 (1984)].
2. A. I. Buzdin and L. N. Bulaevskiĭ, Usp. Fiz. Nauk **149**, 45 (1986) [Sov. Phys. Usp. **29**, 412 (1986)].
3. Yu. A. Izyumov, N. M. Plakida, and Yu. N. Skryabin, Usp. Fiz. Nauk **159**, 621 (1989) [Sov. Phys. Usp. **32**, 1060 (1989)].
4. *Superconductivity in Ternary Compounds*, Vol. 2: *Superconductivity and Magnetism*, Ed. by E. Fisher and M. Maple (Springer, Heidelberg, 1982; Mir, Moscow, 1985).
5. *Physical Properties of High-Temperature Superconductors*, Ed. by D. M. Ginsberg (World Sci., Singapore, 1989; Mir, Moscow, 1990), Chaps. 4 and 6.
6. V. I. Karpman, *Nonlinear Waves in Dispersive Media* (Nauka, Moscow, 1973).
7. B. B. Kadomtsev, *Collective Phenomena in Plasmas* (Nauka, Moscow, 1976).
8. G. A. Askar'yan, Zh. Éksp. Teor. Fiz. **42**, 1567 (1962) [Sov. Phys. JETP **15**, 1088 (1962)]; V. I. Bespalov and V. I. Talanov, Pis'ma Zh. Éksp. Teor. Fiz. **3**, 471 (1966) [JETP Lett. **3**, 307 (1966)].
9. M. J. Lighthill, J. Inst. Math. Appl. **1**, 269 (1965).
10. A. Hasegawa, Opt. Lett. **9**, 288 (1984).
11. N. N. Akhmediev, V. M. Eleonsky, and N. E. Kulagin, Izv. Vyssh. Uchebn. Zaved. Radiofiz. **31**, 244 (1988).
12. N. M. Ercolani, M. G. Forest, and D. W. McLaughlin, Appl. Math. Lett. **23**, 149 (1986).
13. M. Islam, *Ultra Fast Optical Devices* (Oxford Univ. Press, Oxford, 1993).
14. M. Alfimov, V. M. Eleonsky, and N. E. Kulagin, Chaos **22**, 454 (1992).

15. Yu. M. Aliev, V. P. Silin, and S. A. Uryupin, Sverkhprovodimost: Fiz. Khim. Tekh. **5**, 228 (1992).
16. A. Gurevich, Phys. Rev. B **46**, 3187 (1992).
17. Yu. M. Ivanchenko and T. K. Soboleva, Pis'ma Zh. Éksp. Teor. Fiz. **51**, 100 (1990) [JETP Lett. **51**, 114 (1990)].
18. Yu. M. Ivanchenko and T. K. Soboleva, Phys. Lett. A **147** (1), 65 (1990).
19. Yu. M. Ivanchenko and T. K. Soboleva, Fiz. Tverd. Tela (Leningrad) **32**, 2029 (1990) [Sov. Phys. Solid State **32**, 1181 (1990)].
20. R. G. Mints and I. B. Snapiro, Phys. Rev. B **51**, 3054 (1995).
21. A. I. Lomtev, Pis'ma Zh. Éksp. Teor. Fiz. **69**, 132 (1999) [JETP Lett. **69**, 148 (1999)].
22. A. I. Lomtev, Fiz. Tverd. Tela (St. Petersburg) **42**, 16 (2000) [Phys. Solid State **42**, 15 (2000)].
23. A. I. Lomtev, Zh. Tekh. Fiz. **70** (9), 63 (2000) [Tech. Phys. **45**, 1159 (2000)].
24. I. O. Kulik and I. K. Yanson, *The Josephson Effect in Superconducting Tunnel Structures* (Nauka, Moscow, 1970).
25. Yu. E. Kuzovlev and A. I. Lomtev, Zh. Éksp. Teor. Fiz. **111**, 1803 (1997) [JETP **84**, 986 (1997)].
26. A. I. Lomtev, Zh. Éksp. Teor. Fiz. **113**, 2256 (1998) [JETP **86**, 1234 (1998)].
27. F. Kh. Abdullaev, Pis'ma Zh. Tekh. Fiz. **23** (2), 8 (1997) [Tech. Phys. Lett. **23**, 52 (1997)].
28. B. I. Halperin and P. C. Hohenberg, Phys. Rev. **188**, 898 (1969).
29. P. M. Richards and M. B. Salamon, Phys. Rev. B **9**, 32 (1974).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1429–1435. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 69–75.

SOLIDS

Localized Strain Autowaves at the Initial Stage of Plastic Flow in Single Crystals

V. I. Danilov, S. A. Barannikova, and L. B. Zuev

Institute of Strength Physics and Materials Science, Siberian Division, Russian Academy of Sciences, Akademicheskiĭ pr. 2/1, Tomsk, 634021 Russia

e-mail: levzuev@mail.tomsknet.ru

Received April 3, 2003

Abstract—Plastic strain localization in single crystals of pure metals and alloys is studied on the yield plateau and at the easy glide stage with a zero or small strain hardening coefficient. The difference between localization patterns in the two cases is explained, and strain localization mechanisms are suggested. At these stages of plastic deformation, various types of autowaves are observed. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

In a number of recent works [1–3], a unique correspondence between the strain hardening law $\Theta = d\sigma/d\varepsilon = \Theta(\varepsilon)$ and the strain localization pattern at various stages of plastic deformation has been established. Detailed analysis of plastic strain patterns has shown [3] the presence of specific localization patterns at $\Theta = 0$, $\Theta = \text{const} \neq 0$, and $\Theta \sim \sqrt{\varepsilon}$.

The evolution of strain autowaves has been comprehensively analyzed under the intense extension of single-crystalline metals and alloys [3–7]. It was found in these works that, for the three-stage stress–strain curve, a solitary zone of localized strain (switching autowave) propagates along the sample at the easy glide stage and on the yield plateau; several equidistant localization zones (phase autowave) move with a constant velocity at the stage of linear hardening; and a set of periodically arranged immobile local strain maxima, between which the material is virtually unstrained (steady-state dissipative structure), arises at the parabolic stage. The absence of any of the stages in the flow curve implies the related deformation type.

In the above scenario (when we deal with clearly defined process stages, such as linear or parabolic), strain pattern identification is straightforward. However, the difference (if any) between the localization types on the yield plateau and at the easy glide stage is sometimes hardly discernible for single crystals, since the strain hardening coefficient in the latter case is close to zero ($0 < \Theta \approx 10^{-4}G$, where G is the shear modulus of a crystal). This makes it difficult to distinguish between the yield plateau and the easy glide stage ($\theta = 0$). The situation is aggravated by the fact that, in a number of materials, the yield plateau follows the easy glide stage, the transition between them being vague [8]. Glide traces may not shed light on the stage of plastic flow especially in the case of heavily doped single crystals [8]. In this work, we perform comparative analysis of strain localization on the yield plateau and the easy glide stage in order to find the difference in the plastic flow stages.

Such a problem stems from the fact that the localization is not a random process. Being spatially and temporally ordered, it is of autowave character. The type of strain autowaves is defined by the dependence of the strain hardening coefficient Θ on the total strain, i.e., on the loading stage in the stress–strain curve [3].

MATERIALS AND INVESTIGATION TECHNIQUES

The objects of investigation were Cu and Ni single crystals, single-crystalline Fe- and Cu-based alloys, and NiTi intermetallic (Table 1). Cu and Ni are deformed by dislocation glide. In Fe-based alloys, one can induce dislocation glide or twinning by varying the concentration of interstitial impurities (C and N) and orientation of the extension axis. NiTi single crystals of the composition given in Table 1 are deformed through the $B2 \longrightarrow B19'$ phase transformation at 300 K.

All the single crystals were grown by the Bridgman method in an inert atmosphere. Samples in the form of double blades were prepared from homogenized ingots by arc cutting. The working surface area of the samples was 28×5 mm (20×5 mm for titanium nickelide), and their width was varied from 1.3 to 1.5 mm (the width of the Cu single crystals was 3 mm). The alloy samples were quenched in cold water after 1-h keeping at the homogenization temperature. The samples thus prepared were extended with an Instron-1185 machine. Simultaneously, the field of displacement vectors $r(x, y)$ for points on the sample surface were recorded by the method of speckle interferometry, starting from the yield point up to rupture [1]. By numerically differentiating these fields with respect to the coordinates x and

1063-7842/03/4811-1429$24.00

Table 1. Chemical composition of the metals and alloys (wt %)

Material	Cu	Ni	Fe	Ti	Cr	Mn	Sn	Mo	C	N
Copper	99.98	–	–	–	–	–	–	–	–	–
Nickel	–	99.98	–	–	–	–	–	–	–	–
Cr–Ni austenite, A_I	–	12.4	65.5	–	18.0	2.3	–	1.2	0.013	0.35
High-manganese austenite, A_{II}	–	–	85.96	–	–	13.0	–	–	0.9–1.0 3	–
Cu–Ni–Sn alloy	84.0	10.0	–	–	–	–	6.0	–	–	–
Titanium nickelide (at. %)	–	50.3	–	49.7	–	–	–	–		

y, the longitudinal, ε_{xx}, transverse, ε_{yy}, shear, $\varepsilon_{xy} = \varepsilon_{yx}$, and rotational, ω_z, components of the plastic strain tensor

$$\nabla r(x, y) = \begin{vmatrix} \varepsilon_{xx} & \varepsilon_{xy} \\ \varepsilon_{yx} & \varepsilon_{yy} \end{vmatrix} + \omega_z$$

[9] for the sample surface was calculated. The spatial, $\varepsilon_{xx}(x, y)$, and spatial–temporal, $\varepsilon_{xx}(x, t)$, distributions of local elongations $\varepsilon_{xx} = du/dx$ (where u is the component of the vector r in the direction of the extension axis x) were analyzed. This allowed us to locate strain localization zones and trace their evolution with time.

The additional microscopic investigation of steps on the surface made it possible to refine the indices of slip and twinning systems.

STRAIN LOCALIZATION DYNAMICS IN CRYSTALS WITH VARIOUS DEFORMATION MECHANISMS

The micro- and macromodels of plastic flow in alloys and pure metals differ markedly. In alloys, usually one slip system is observed even in samples oriented for multiple slip. This causes the easy slip stage and sometimes the appearance of Lüders bands. Therefore, the extension of all the single crystals gave a set of stress–strain diagrams the initial portion of which had a zero or a very small slope (strain hardening coefficient Θ).

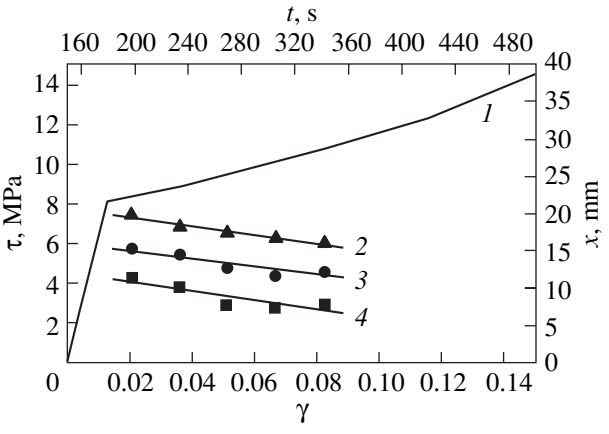


Fig. 1. Initial stages of the plastic flow curve for the copper single crystal and the space–time evolution of localized strain zones at stage I.

Consider strain localization in pure single-crystalline metals. The extension axis of Cu samples was oriented along the direction $[\bar{1}39]$, and the surface under study had indices $(3\bar{2}1)$. With such an orientation, the $(111)[\bar{1}01]$ slip system appeared once the yield point was reached. The strain rate was 5.5×10^{-5} s^{-1}. The strain fields were recorded during loading in 36-s intervals (the total strain increment was $\delta\varepsilon_{tot} = 0.2\%$).

The stress–strain curves obtained under such conditions had three stages [10, 11]. Figure 1 (curve *1*) shows the part of the loading diagram corresponding to the easy slip stage (I), a part of the linear stage (II), and the transition zone between them. The parameters of plastic flow, namely, the critical shear stress τ_0, the strain γ_1 at the end of the easy slip stage, and the strain hardening coefficients at the easy slip stage (Θ_I) and linear stage (Θ_{II}), are listed in Table 2. For the crystals with the given orientations, they are close to those reported in [10, 11].

The distribution of local elongations at the easy slip stage of the Cu single crystals had the form of three wide equidistant zones moving synchronously along the sample (Fig. 2). Curves *2–4* demonstrate the variation of the positions of these zones with time. The strain localization zone velocity estimated from the slope to the *t* axis was found to be $\approx 5.3 \times 10^{-5}$ m/s.

The single-crystalline Ni samples with the surface $(0\bar{7}6)$ were extended in the direction $[\bar{1}67]$. With this orientation, the slip system $(111)[\bar{1}01]$ was observed. The strain rate and the interval over which the strain fields were recorded were the same as for the Cu samples. The flow curves for Ni also have three stages. The initial portion of a typical curve (curve *1*) is shown in Fig. 3. The parameters of the loading curves for Ni (Table 2) also agree with [10, 11].

At the easy slip stage (I), two deformation fronts move toward each other. Their positions during deformation are shown in Fig. 3 (curves *2*, *3*). The velocities of the fronts were $+3.6 \times 10^{-5}$ and -2.2×10^{-5} m/s (the plus and minus signs mean that the fronts move in opposite directions). The point of their meeting corresponds to the end of stage I.

In single-crystalline Cu- and Fe-based alloys, which are deformed by slipping, flow curves and strain localization patterns are more diversified. The Cu–Ni–Sn alloy were quenched in water starting from 1203 K in order to fix the homogeneous solid solution of Ni and Sn in Cu. The extension axis was aligned with the $[1\bar{1}1]$ direction, and the surface had indices $(\bar{3}58)$. In this case, the three slip systems $(111)[0\bar{1}1]$, $(11\bar{1})[011]$, and $(\bar{1}11)[0\bar{1}1]$ are equivalent, and starting from the yield point, one might expect multiple slip. However, when these samples were extended with a rate of 6.6×10^{-5} s^{-1}, the $(111)[0\bar{1}1]$ slip system alone was observed and the strain–stress curve (Fig. 4, curve *1*) exhibited both the sharp yield point and yield plateau (stage I) followed by the extended stage of linear hardening (stage II). The material hardening parameters are listed in Table 2. The localized strain field was recorded in 0.2% increments of the total strain in the interval from 0.7 to 7.3%. Immediately after the appearance of the sharp yield point, a solitary strain localization front is observed near the movable tension grip. This front travels along the sample throughout the yield plateau (Fig. 4, curve *2*) with a velocity of $\approx 6.7 \times 10^{-5}$ m/s.

The samples of Cr–Ni austenite are also deformed by dislocation glide. They were saturated by nitrogen with the thermobaric method at 1473 K, kept in a vacuum for 1 h, and then quenched in water in order to fix the homogeneous substitutional Fe-based solid solution. Nitrogen in this solid solution acts as an interstitial impurity without forming nitrides. The samples with the (110) surface were extended along the direction [111]. With such an orientation, three slip systems, $(11\bar{1})[101]$, $(1\bar{1}1)[011]$, and $(\bar{1}11)[\bar{1}0\bar{1}]$, have a Schmid factor of 0.27. However, the presence of the overequilibrium interstitial hardening impurity leads to a substantial increase in the deforming stress and the degree of dissociation of dislocations. The latter fact, in turn, restricts slip localization to the system $(1\bar{1}1)[011]$, which dominates over the other two at the early stage of deformation. The strain rate and the time instants of strain recording were the same as for Cu and Ni. Under these conditions, the loading curve also has three stages, as in the case of Cu and Ni single crystals which were oriented for single slip, but, unlike Cu and Ni, has a sharp yield point.

Figure 5 (curve *1*) shows the part of the stress–strain diagram covering the easy slip portion (I), transition region, and the beginning of the linear stage (II). The basic parameters of the hardening curve are listed in Table 2. Once the sharp yield point had been reached, deformation proceeded with a small but nonzero hardening coefficient (Table 2). The positions of the local strain peaks are shown by curves *2–4* in Fig. 5. Up to the value of the shear strain $\gamma \approx 0.04$, the local elongation distribution was random; then, three strain zones formed. One traveled with a constant velocity of $\approx 3.5 \times 10^{-5}$ m/s throughout the easy slip stage; the other two first moved faster than the first zone but then stopped. During stage I, the entire sample was covered by these moving strain localization zones. The strain zone passed along each section of the sample once.

Table 2. Characteristics of plastic flow in single crystals deformed by dislocation glide

Material	τ_0, MPa	$\Theta_I/G \times 10^{-4}$	$\Theta_{II}/G \times 10^{-3}$	γ_1
Cu	6.2	7.5	2.6	0.14
Ni	23.0	11	2	0.06
Cu–Ni–Sn	44.8	0	1.1	0.037
A_I	76.0	2.3	1.2	0.15

The homogenization of high-manganese austenitic steel samples with a carbon content of ≈1 wt%, which were deformed by twinning, was performed by keeping

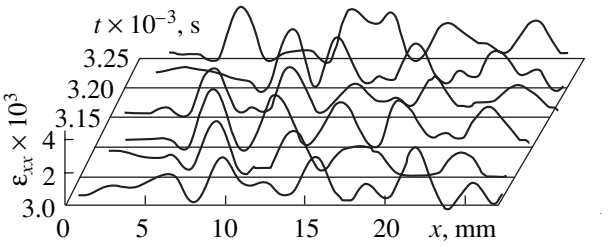


Fig. 2. Distribution of the localized strain zones in the copper single crystal at the easy slip stage (stage I).

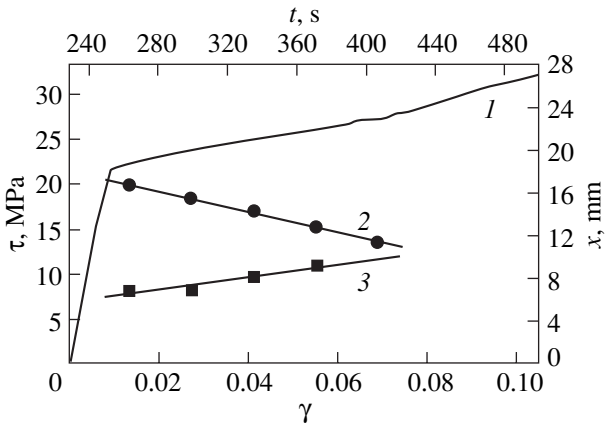


Fig. 3. Initial stages of the plastic flow curve for the nickel single crystal and the space–time evolution of localized strain zones at stage I.

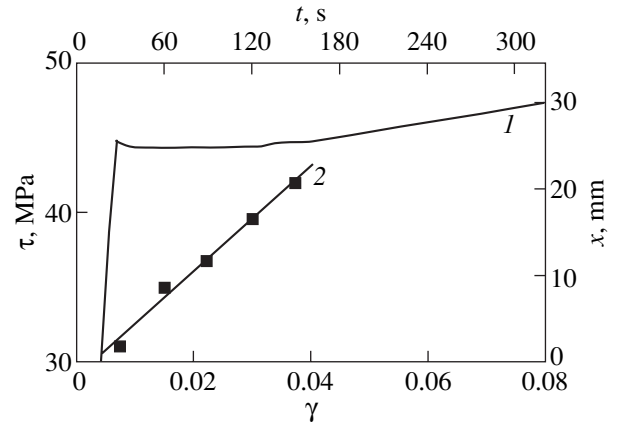


Fig. 4. Initial stages of the plastic flow curve for the Cu–Ni–Sn single crystal and the motion of the localized strain zone at stage I.

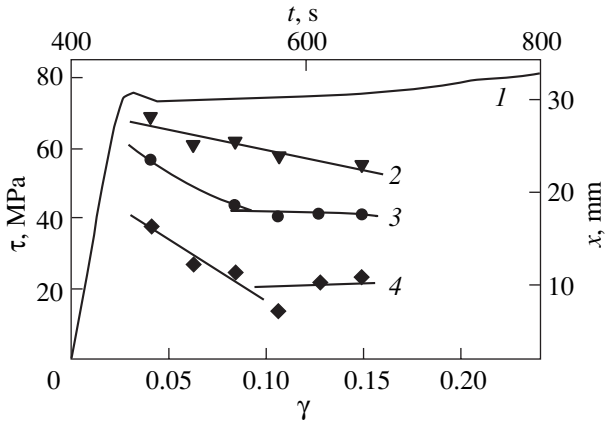


Fig. 5. Initial stages of the plastic flow curve for the single crystal A_I and the space–time evolution of localized strain zones at stage I.

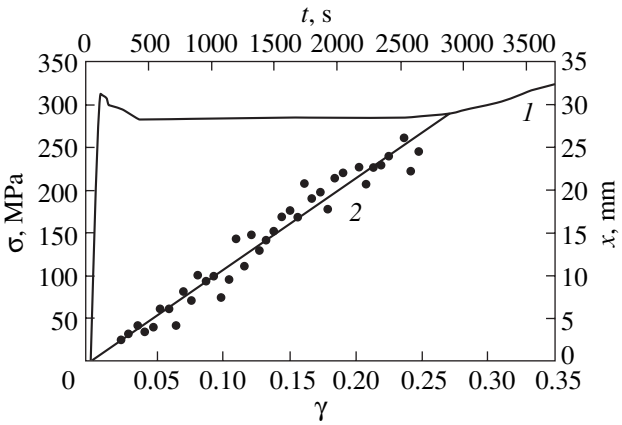


Fig. 6. Initial stages of the plastic flow curve for the single crystal A_{II} (the extension axis $[\bar{3}77]$) and the motion of localized strain zones at stage I.

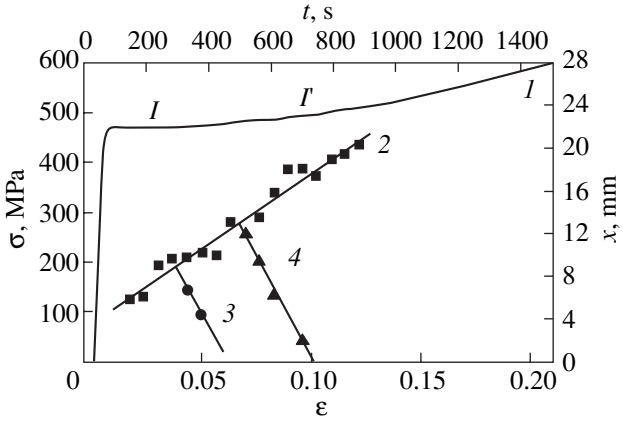


Fig. 7. Initial stages of the plastic flow curve for the single crystal A_{II} (the extension axis $[\bar{3}55]$) and the space–time evolution of localized strain zones at stages I and I'.

at 1373 K for 1 h. Then they were quenched in water. The extension axis had two directions: $[\bar{3}77]$ and $[\bar{3}55]$. The working surface of the samples was (011) in both cases. The samples were extended at room temperature with a strain rate of 1.2×10^{-4} s^{-1}. Starting from the yield stress, the displacement fields were recorded every 15 s (in 0.2% increments of the total strain) or every 36 s (in 0.4% increments).

In the single crystals oriented along $[\bar{3}77]$, the basic plastic deformation mechanism under room-temperature extension is twinning in the system $(111)[\bar{2}11]$ [8], which is observed after the yield point. The length of the yield plateau (stage I) together with the sharp yield point was about 30%. Subsequent loading leads to the stage of linear hardening (stage II), which extends up to rupture at ≈45% of the total strain. In single crystals of such an orientation, the parabolic hardening stage is absent.

Figure 6 (curve *1*) shows the part of the strain–stress diagram for the high-manganese austenitic single crystal with the extension axis $[\bar{3}77]$. This part of the diagram covers stage I and a part of stage II. Since plastic flow in this case is due to twinning, the diagram is plotted in the σ–ε, rather than τ–γ, coordinates. The hardening parameters are listed in Table 3. In this case, the plastic strain front travels along the sample throughout the yield plateau (stage I in the strain–stress curve) with a constant velocity of $\approx 9 \times 10^{-6}$ m/s (Fig. 6, curve *2*).

The strain pattern on the samples with the extension axis $[\bar{3}55]$ is more complicated. It is known [4] that, when the C content in single-crystalline Hadfield steel is ≈1 wt% and the extension axis is $[\bar{1}11]$, multiple twinning in the systems $(111)[\bar{2}11]$, $(\bar{1}\bar{1}1)[\bar{1}21]$, and $(1\bar{1}1)[\bar{1}12]$ takes place and the loading diagram has only the linear hardening stage up to rupture. The direc-

tion (orientation) [$\bar{3}55$] falls into the standard stereographic triangle and lies between two extreme cases: [$\bar{1}11$] and [$\bar{3}77$]. Therefore, the strain–stress curve for such samples has both the sharp yield point and yield plateau (stage I) before the linear hardening stage and plastic deformation starts from twinning in the system (111)[$\bar{2}11$].

The transition to multiple twinning and to linear stage II, which has a high hardening coefficient Θ_{II}, involves one more linear portion (stage I') with a much lower hardening coefficient $\Theta_{I'}$ (Table 3; Fig. 7, curve *1*). The analysis of the local strain distribution indicated that a solitary strain zone travels from the fixed tension grip with a velocity of $\approx 1.9 \times 10^{-5}$ m/s on the yield plateau (stage I). The variation of the position of this localization zone is depicted in Fig. 7 (curve *2*). At stage I', another front separates from the primary strain zone (Fig. 7, curve *3*). The primary strain localization zone continues to move along the unstrained part of the sample with the same velocity, while the new one travels over the strained part with a velocity of -5.5×10^{-5} m/s. Such a situation occurs at stage I' twice (Fig. 7; curves *3*, *4*). Stage I' is complete when the primary front passes through the entire sample (Fig. 7).

The composition of titanium nickelide single crystals (Table 1) provides the smooth phase transition $B2 \longrightarrow B19'$ [12] at 300 K. The extension axis of the (110) sample was aligned with the direction [$\bar{1}12$] in the *B*2 lattice. The strain rate was 8.3×10^{-5} s^{-1}. Under these conditions, the strain–stress curves exhibit a sharp yield point and a yield plateau several percent long. The latter is followed by a short transition region and the parabolic stage of strain hardening. As is known [12], it is the phase transition $B2 \longrightarrow B19'$ that is responsible for a yield plateau and the parabolic stage corresponds to the deformation of already formed martensite.

Figure 8 (curve *1*) shows the part of the σ–ε diagram for the TiNi sample covering the sharp yield point and yield plateau (stage I) and the transition region. The plastic flow parameters are given in Table 3. From the analysis of the local strain distribution, it follows that a strain front ≈2 mm wide travels along the sample with a constant velocity of $\approx 3 \times 10^{-5}$ m/s during the yield plateau. The variation of its position in the course of deformation is shown in Fig. 8 (curve *2*). As in all the cases described above, the yield plateau is complete when the localized strain front passes through the whole sample.

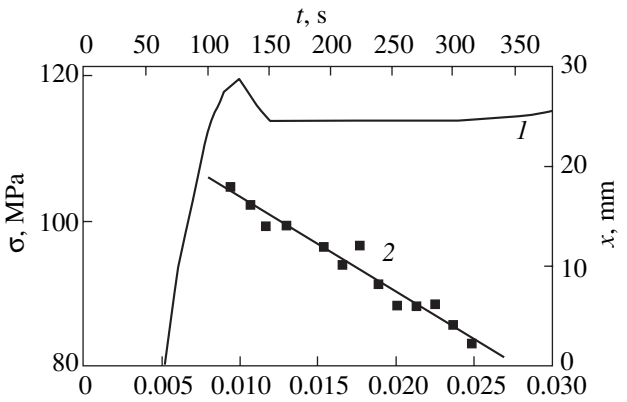


Fig. 8. Initial stages of the plastic flow curve for the NiTi single crystal and the motion of localized strain zones at stage I.

RESULTS AND DISCUSSION

The analysis of the results obtained in this paper allows us to establish a qualitative correlation between the initial stages of the plastic flow curve for single-crystalline materials and the space–time local strain distributions. The basic conclusions are as follows.

(1) If the flow curve of a single crystal has a sharp yield point and a yield plateau ($\Theta_I = 0$), a traveling solitary front separating elastically and plastically strained regions is observed (Cu–Ni–Sn, Fig. 4; Mn austenite with the extension axis aligned with the [$\bar{3}77$] direction, Fig. 6, and TiNi, Fig. 8) irrespective of the composition, crystal structure, and orientation of the material, as well as of plastic flow micromechanisms.

(2) If the stress–strain curve exhibits the linear stage with a small but nonzero hardening coefficient (Θ_I = const > 0, easy glide in fcc single crystals with the dislocation mechanism of deformation), several traveling strain localization zones are observed in all the samples studied (Figs. 1, 3, 5). They may move in the same or opposite directions, and their velocities may differ or equal each other.

(3) If the yield plateau is followed by the linear stage with a small strain hardening coefficient, secondary

Table 3. Characteristics of plastic flow in single crystals with nondislocation mechanisms of deformation

Material	σ_0, MPa	$\Theta_I/G \times 10^{-4}$	$\Theta_{I'}/G \times 10^{-4}$	$\Theta_{II}/G \times 10^{-3}$	ε_I	$\varepsilon_{I'}$
A_{II}, [$\bar{3}$ 77]	314.5	0	–	6.6	0.25	–
A_{II}, [$\bar{3}$ 55]	469.4	0	40	12.9	0.04	0.1
NiTi	119.5	0	–	–	0.03	–

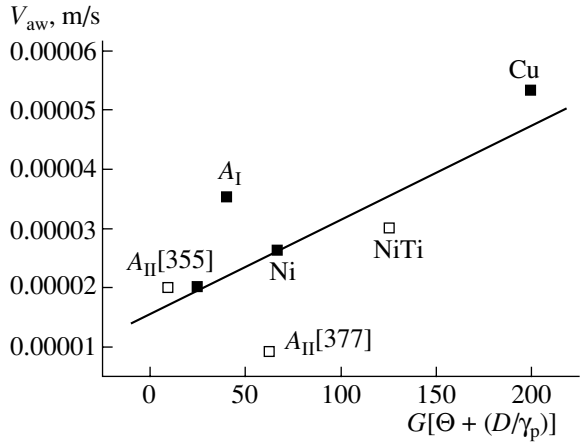


Fig. 9. Localized strain autowave velocity vs. reduced hardening coefficient for the single crystals at stage I. □, yield plateau; ■, easy slip.

fronts separate out from the traveling primary strain zone. They move in the opposite direction over the plastically strained part of the sample (Mn austenite with the $[\bar{3}55]$ extension axis, Fig. 7).

In the first case of localization, the traveling front originating at the upper yield point turns the material from one state into another (Lüders band [13]). This is clearly observed in TiNi, where the phase transition $B2 \longrightarrow B19'$ occurs at the deformation front. Such a front may pass through the material only once. If several such fronts originate, they cancel each other out when meeting. The flow proceeds without hardening until all elements of a deforming volume become strained.

In the second case of localization, several strain zones travel along the sample at stage I. They originate prior to this stage from randomly distributed localization zones. In pure metals (Cu and Ni), the random distribution corresponds to the smooth transition from elasticity to fully developed plasticity; in concentrated solid solutions additionally strengthened by interstitial impurities (Cr–Ni austenite), to a sharp yield point. Hence, the strain fronts become regularly arranged and move in the deformed, though slightly, material. Of most interest is the situation when the linear stage with a small hardening coefficient (stage I') follows the yield plateau (Mn austenite with the $[\bar{3}55]$ extension axis). Here, secondary strain localization zones move oppositely to the primary front. It should be emphasized that each of the strain localization maxima (fronts) passes through the sample once. This is clearly seen in Figs. 5 (curves *2–4*) and 7 (curves *3*, *4*).

The knowledge of the velocity of strain localization zones at the material loading stages considered is important for the elaboration of an autowave model of plastic flow. It has been shown [3] that the velocity V_{aw} of plastic strain autowaves at the easy slip stage in a number of single crystals is given by

$$V_{aw} = V_\infty + \xi\frac{G}{\Theta}. \quad (1)$$

While the plastic flow localization patterns on the yield plateau and at the easy slip stage are similar, formula (1) yields $V_{aw} \longrightarrow \infty$ with $\Theta = 0$, which is typical only of the yield plateau. To improve the situation, one should take into account the effective friction stress D of mobile dislocations at the yield stress. According to Gilman [14], this effective stress controls the mobility of dislocations:

$$V_{disl} = V^* \exp\left(-\frac{D}{\tau}\right).$$

Following Gilman [14], the denominator of the second term in Eq. (1) is replaced by $\Theta + D/\gamma_p$:

$$V_{aw0} = V_0 + \xi\frac{G}{\Theta + D/\gamma_p}. \quad (2)$$

With $D \approx 2\tau_0$ [14], $\tau_0 \approx 2.5 \times 10^{-5}G$ for most fcc single crystals, and $5 \times 10^{-3} < \gamma_p < 5 \times 10^{-2}$ [8] at the easy slip stage, we may estimate the value of V_{aw0}, which is the velocity of autowaves at $\Theta = 0$.

The results of processing the data for the front velocities at the easy slip stage and on the yield plateau that are estimated by formula (2) are summarized in Fig. 9. The correlation factor for the variables is ≈0.75, that is, significant [15]. In the case of linear hardening, $\Theta \gg D/\gamma_p$ and the conventional relationship $V_{aw} \sim 1/\Theta$ [7] is valid.

CONCLUSIONS

Thus, the type of localized strain autowaves at the initial stages of plastic flow in single crystals depends on a relationship between the stress applied and the total strain. With $\Theta_I = 0$, a switching autowave in the form of a strain localization zone is generated. When traveling along the sample, this wave switches it to another state. In this case, stage I is complete when the entire sample passes to a new plastically strained state. When $\Theta_I = \text{const} > 0$, an autowave comprises several strain localization zones which travel in already strained volume. Each of the strain localization maxima (zones) passes through a certain part of the sample only once. This suggests the propagation of an excitation autowave.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education of the Russian Federation (grant no. PD 02-1.2-63).

REFERENCES

1. L. B. Zuev and V. I. Danilov, Philos. Mag. A **79** (1), 43 (1999).
2. L. B. Zuev, Annalen der Physik **10**, 965 (2001).
3. L. B. Zuev, V. I. Danilov, and B. S. Semukhin, Usp. Fiz. Met. (Ukr.) **3**, 237 (2002).
4. L. B. Zuev, S. A. Barannikova, V. I. Danilov, *et al.*, Zh. Tekh. Fiz. **69** (10), 56 (1999) [Tech. Phys. **44**, 1179 (1999)].
5. L. B. Zuev, V. I. Danilov, and S. A. Barannikova, Int. J. Plast. **17** (1), 47 (2001).
6. S. A. Barannikova, V. I. Danilov, and L. B. Zuev, Zh. Tekh. Fiz. **72** (9), 63 (2002) [Tech. Phys. **47**, 1130 (2002)].
7. V. I. Danilov, S. A. Barannikova, K. V. Gonchikov, and L. B. Zuev, Kristallografiya **47**, 730 (2002) [Crystallogr. Rep. **47**, 672 (2002)].
8. I. Karaman, H. Sehitoglu, Yu. I. Chumlyakov, *et al.*, Metall. Mater. Trans. A **32**, 695 (2001).
9. A. Kadich and D. G. B. Edelen, *A Gauge Theory of Dislocations and Disclinations* (Springer-Verlag, Berlin, 1983; Mir, Moscow, 1987).
10. F. R. N. Nabarro, Z. S. Bazinskiĭ, and D. B. Kholt, *Plasticity of Pure Single Crystals* (Metallurgiya, Moscow, 1967).
11. R. Berner and H. Kronmüller, *Plastische Verformung von Einkristallen*, in *Moderne Probleme der Metallphysik*, Ed. by A. Seeger (Springer-Verlag, Berlin, 1965; Mir, Moscow, 1969).
12. S. Miyazaki, S. Kimura, F. Takei, *et al.*, Scr. Metall. **17**, 1057 (1983).
13. D. McLean, *Mechanical Properties of Metals* (Wiley, New York, 1963; Metallurgiya, Moscow, 1965).
14. J. J. Gilman, J. Appl. Phys. **36**, 2772 (1965).
15. A. K. Mitropol'skiĭ, *Techniques for Statistical Calculation* (GIFML, Moscow, 1961).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1436–1441. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 76–82.

SOLID-STATE ELECTRONICS

Synthesis of Single-Walled Carbon Nanotubes in an Expanding Vapor–Gas Flow Produced by Laser Ablation of a Graphite–Catalyst Mixture

G. I. Kozlov* and I. G. Assovskiĭ**

** Institute of Problems of Mechanics, Russian Academy of Sciences, Moscow, 119526 Russia*
e-mail: kozlov@ipmnet.ru
*** Semenov Institute of Chemical Physics, Russian Academy of Sciences, ul. Kosygina 4, Moscow, 119991 Russia*

Received December 15, 2002

Abstract—Single-walled carbon nanotubes (SWCNTs) are synthesized by the ablation of a catalyst-containing carbon target with a cw CO_2 laser. Emphasis is on ablation conditions that are favorable to self-organized SWCNT synthesis. It is shown that the graphite target intensely evaporates with the formation of fractal-like tubes at the edge of the jet when the laser power density exceeds 10^5 W/cm^2. Still more favorable conditions for carbon nanotube synthesis are set if the power density lies within 2×10^4–5×10^4 W/cm^2. Under these conditions, both individual SWCNTs and their bundles of diameter from 1.1 to 1.5 nm are produced, as shown by Raman scattering and electron microscopy studies. In this series of experiments, the maximal fraction of SWCNTs reaches 20%. A mechanism of SWCNT fast growth in the laser torch is suggested. © *2003 MAIK "Nauka/Interperiodica".*

The search for effective procedures for the synthesis of carbon nanostructures, including nanotubes [1], and the elucidation of underlying mechanisms is a very important scientific problem of current interest.

A perfect SWCNT is known to be a jointless graphene sheet folded into a cylinder with a wall thickness on the order of the carbon atom size. Experiments on SWCNT synthesis by using an electric arc, pyrolysis of hydrocarbons, and laser ablation of carbon [2] (a review of SWCNT synthesis methods is given elsewhere [3, 4]) suggest that carbon nanotubes can be synthesized by low-temperature methods, which involve mostly the pyrolysis of hydrocarbons, and by high-temperature ones (where the temperature exceeds the upper limit of graphite thermal stability, 2600 K), which are associated with the evaporation of a catalyst-containing graphite target. The latter are of special significance because the process goes under more refined conditions with the participation of carbon and inductor (catalyst) vapors in the molecular state. This provides the possibility of tracing the mechanism and kinetics of subsequent processes: the condensation of vapor and the self-organized nucleation and growth of molecular associates, compact clusters, and carbon nanostructures.

In this paper, we report the results of pioneering experiments in which laser ablation conditions provided the self-organized synthesis of carbon nanotubes in the expanding flux of a carbon–catalyst target. We were interested in irradiation power densities such that the temperature of an absorbing target layer does not exceed the critical value and the solid–liquid–vapor phase transition takes place.

LASER SETUP FOR NANOSTRUCTURE SYNTHESIS

The laser ablation of carbon nanotubes was accomplished on the modernized setup (Fig. 1) that was applied earlier in experiments on the propagation of pyrolysis wave along a laser beam [5]. The radiation of a 2-kW cw CO_2 laser with a beam divergence of 8×10^{-4} rad was focused by a KCl lens with a focal length of 20 cm, passed through a salt window inside the 51-cm-long quartz reaction tube of inner diameter 7.1 cm, and was concentrated on the end face of a cylindrical graphite target. With a laser beam density of 1–2 kW, the target temperature falls into the interval 3500–

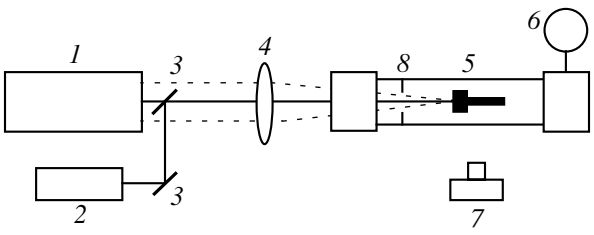


Fig. 1. Experimental setup for laser synthesis of carbon nanotubes: (*1*) CO_2 laser, (*2*) He–Ne laser, (*3*) reflectors, (*4*) lens, (*5*) reactor, (*6*) graphite target, (*7*) optical pyrometer, and (*8*) diaphragm.

1063-7842/03/4811-1436$24.00 © 2003 MAIK "Nauka/Interperiodica"

4000 K, where radiation losses prevail, if the target diameter is smaller than 5 mm, as follows from calculations. Therefore, the diameter of the target, which was flowed about by helium at a pressure of 400–760 torr, was varied between 3 and 5 mm in our experiments. The helium flow rate through the reaction tube was kept at a level of 0.5 l/min. Helium was used because it is a good relaxant of the vibrational energy of the resulting nanostructures. The diameter of the laser spot on the end face of the target was 2 mm. The brightness temperature of the target surface was monitored with an optical pyrometer.

The laser radiation enhanced the target vaporization, forming an ablation product jet directed toward the laser beam and helium flow. The optics at the entrance was protected against this jet with a special diaphragm, which was mounted inside the reaction tube. The diaphragm turned the ablation product jet, directing it along the wall toward the exit from the reactor. Particles formed in the peripheral (colder) regions of the jet condensed on the diaphragm, reactor walls, and special substrates placed throughout the reactor length. Substrate deposits were examined by Raman spectroscopy [6], as well as by (transmission and scanning) electron microscopy combined with the JEOL JSM-5410 + XLink Oxford Detector X-ray microanalyzer.

LASER ABLATION REGIMES

The laser power density is the basic factor controlling the process of material ablation. Carboniferous targets exposed to laser radiation are usually porous. Therefore, laser radiation first disperses and then heats the target, forming a polydisperse particle flux, which evaporates in the laser beam. Yet, the radiation–target interaction is an essentially steady process, although the flux density of target vaporization products is significant. One might expect that soot and metal particles contained in target vaporization products shield the radiation. However, this does not happen because the particles in the laser beam have time to heat up to the boiling point and sublimate. Due to the subsequent expansion of the vapor–gas jet and also its mixing with the cold helium flow, the jet cools down and the vapor becomes saturated at some moment. Then, the vapor condenses and clusters start to form. The number and size of condensate nanoparticles depend on the condensation kinetics during the expansion and cooling of the jet. The finely dispersed condensate deposits on the diaphragm, cold walls of the reactor (due to thermophoresis), and substrates. The diaphragm was made of stainless steel, and the substrates had the form of 1-cm-high metallic (copper or brass) cylinders with a diameter of 1 cm and grids made of stainless steel and Nichrome wire.

The morphology of the deposits was found to significantly depend on the optical radiation power (intensity), all other things being equal. With a high beam power density (1.5–2.0 kW) and a small target diameter (3 mm), the target was rapidly heated to temperatures above 3500 K and, accordingly, the carbon vapor density increased substantially. This, in turn, causes the density of compact clusters, which form at the periphery of the vapor–gas jet upon condensation, to grow, and the clusters start to aggregate into fractal microstructures. These structures were detected on all the substrates and in the circumference of the diaphragm.

As the laser radiation threshold intensity for the formation of fractal structures, one may take that providing the intense evaporation of the graphite target, or, in other words, that at which a crater on the target appears fairly rapidly. In our experiments, the crater depth reached about 5 mm for 10 s of irradiation. One could observe that, with the laser power remaining the same, the target heats up further as the crater increases in diameter presumably until the triple point is achieved and the carbon melts.

According to [7], the triple point of carbon is achieved at $p_t = 1$ atm and $T_t = 4000$ K. These values merit special attention, since they imply that the liquid phase of carbon is difficult to obtain at a close-to-normal pressure. We observed the traces of the liquid phase in the form of the solidified flow with characteristic steel color in the throat of and immediately at the exit from the crater. The latter case is depicted in Fig. 2. One can also judge the formation of the carbon liquid phase in the crater from a number of solidified spheroidal droplets near the crater's outlet channel on the end face of the target (see also Fig. 2).

It seems reasonable to examine the structure of the solidified liquid phase of carbon. Without going into detail, we only note that valuable data on the solidified carbon structure may be extracted from Raman spectra. Such a spectrum is shown in Fig. 3 in comparison with the spectrum of the carbon deposit on the diaphragm. It follows that Raman spectra of carbon generally have

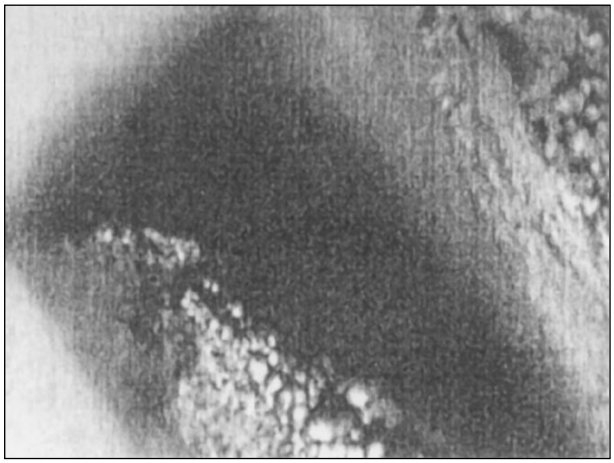


Fig. 2. Micrograph of the solidified liquid phase of carbon at the exit from the crater (×25).

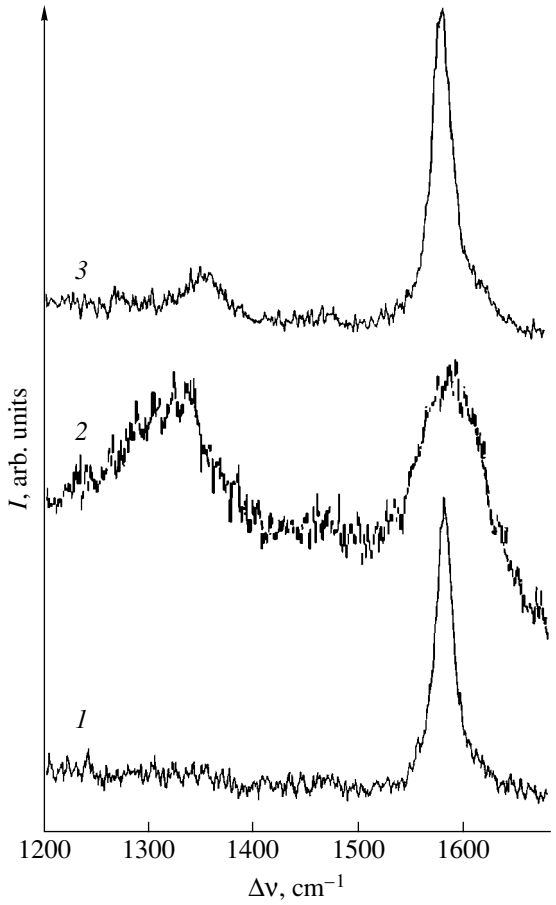


Fig. 3. Raman scattering spectral intensity I vs. shift $\Delta\nu$ of the Raman spectral line: (*1*) solidified liquid phase of carbon, (*2*) carbon deposit on the diaphragm, and (*3*) deposit in the form of fibrous structure on the stainless steel grid.

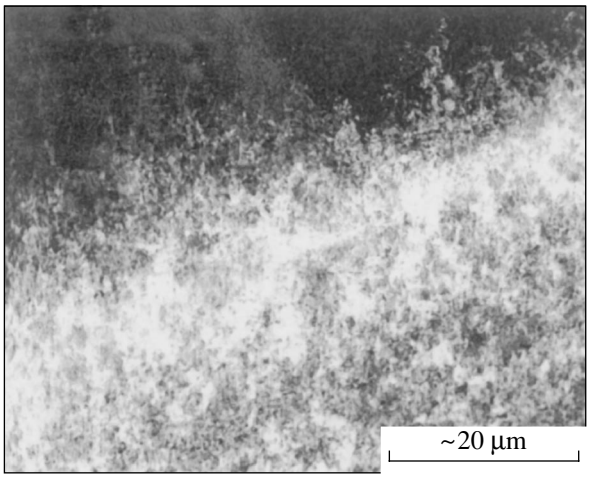


Fig. 4. SEM image of the carbon fibrous deposits on the stainless steel grid.

two modes: G (line shift $\Delta\nu = 1580$ cm^{-1}) and D ($\Delta\nu = 1350$ cm^{-1}). Their widths and relative intensities bear information on the carbon structure.

In the spectrum of the solidified liquid phase, the mode D is totally absent and the mode G is narrow. The latter fact indicates that the structure is highly ordered. So, it seems quite plausible that the carbon solidifies to form either carbon crystals or a crystalline carbyne–graphite mixture, since it is known [8] that, at temperatures above 2600 K, carbon turns into carbyne. This compound has chain molecules, which basically may also play a certain role in nanotube synthesis.

We believe that the aforementioned evaporation conditions with the formation of fractal clusters are inefficient in terms of SWCNT synthesis. Therefore, the laser power was decreased to 0.5–1.5 kW and the diameter of the target was increased to 5 mm. In this way, the laser power density was decreased to $(2–5) \times 10^4$ W/cm^2; accordingly, the evaporation became less intense. Under these conditions, the deposit on the substrate was rather dense but fractals did not form for the irradiation time (10 s). The deposits had a fibrous structure, as follows from Fig. 4, which shows the SEM image of the deposit on the stainless steel grid. This deposit was obtained when the 800-W laser radiation evaporated a purely graphite target in the helium atmosphere under a pressure of 760 torr. In this case, the Raman spectrum (Fig. 3) is typical of a deposit with doubled-walled and multiwalled nanotubes (the G mode is narrow and the D mode is of small intensity).

SYNTHESIS OF SINGLE-WALLED CARBON NANOTUBES

In order to provide more favorable synthesis conditions, a 3.5-mm-diam. channel was made in the graphite target on the irradiated side. The channel was filled with a graphite : nickel catalyst = 19 : 1 mixture. The catalyst was composed of nickel powder and Y_2O_3 additive. The Raman spectra of the deposits on the substrates placed in different parts of the reactor showed that nearly all of them contain single-walled carbon nanotubes.

It should be noted that the rate of target evaporation in these experiments was roughly 10 g/h, which is 50 times higher than the evaporation rate used in [2]. Such a high evaporation rate might adversely affect the growth of nanotubes. The final result was therefore very surprising: the SWCNT maximal concentration was observed on the substrate nearest (2.0 cm) to the target. This means that nanotubes in the laser torch are produced very rapidly and also indicates that the part of the torch adjacent to the evaporation zone on the target has a decisive effect on the nucleation and growth of single-walled nanotubes; hence, the reactor dimensions may be minimized.

A typical Raman spectrum taken from the deposit obtained in these experiments is shown in Fig. 5. Note that the analysis of carbon nanotubes with Raman spectroscopy is a well-developed technique [6]. Its basic advantage is that the presence of nanotubes in the deposits and their characteristics may be established

immediately during the experiments. As follows from Fig. 5, the SWCNT Raman spectrum in this case consists of two modes: tangential and radial. The absence of the *D* mode ($\Delta\nu = 1350$ cm^{-1}) means the absence of the amorphous phase in the deposit. The tangential mode ($\Delta\nu = 1580$ cm^{-1}) has a split typical of single-walled nanotubes. Basically, from the temperature shift of this mode, one can find the SWCNT concentration in the ablation products. The associated results obtained in two experiments with laser radiation powers of 600 and 800 W are shown in Fig. 6, which demonstrates that, all other things being equal, the SWCNT concentration depends markedly on the radiation power under normal helium pressure. It also follows from Fig. 6 that the increase in the laser power by as little as 200 W results in the fourfold increase in the SWCNT concentration (20 vol.% at 800 W).

The radial (or, otherwise, breathing) mode occupies the range $\Delta\nu = 150$–200 cm^{-1} and contains important information on the diameters of SWCNTs in the deposit. The SWCNT diameter *d* may be estimated by the frequency shift of the radial mode from the relationship [9]

$$d = 223.75/\Delta\nu,$$

where $\Delta\nu$ is the frequency shift in inverse centimeters and *d* is the SWCNT diameter in centimeters. From the radial mode spectrum, which represents a set of peaks (Fig. 5), one can estimate the SWCNT diameter distribution. With a radiation power density on the target surface of 5×10^4 W/cm^2, which was used in our experiments, synthesized single-walled nanotubes had a diameter from 1.1 to 1.5 nm. The maximum in this distribution lies at 1.2 nm; thus, it follows that nanotubes with a chirality index of 9.9 prevail in the deposits. In the high-frequency part of the Raman spectrum, the position of the fundamental peak of the tangential mode lies at 1587 cm^{-1}, which also indicates the presence of small-diameter SWCNTs in the deposit [10].

The structure of carbon nanotubes produced by self-organized synthesis may be clarified by taking HRTEM images of the deposit (Fig. 7). They show that individual single-walled nanotubes combine into bundles (with several tens of nanotubes per bundle in our experiments). The diameter of individual nanotubes, which can be estimated from this image, correlates well with its value obtained from the Raman spectrum. It is remarkable that both single and paired nanotubes are rather pure, while the surface of the bundles contains spheroidal nanoparticles of diameter 1–2 nm, which apparently consist of amorphous graphite or catalyst. The appearance of these particles may be associated with the nonuniform distribution of graphite powder grains and catalyst in the target.

To conclude this section, we observed an interesting effect: the formation of extremely thin carbon filaments, which form a weblike macrostructure. This occurred, as a rule, upon grid condensation, as well as

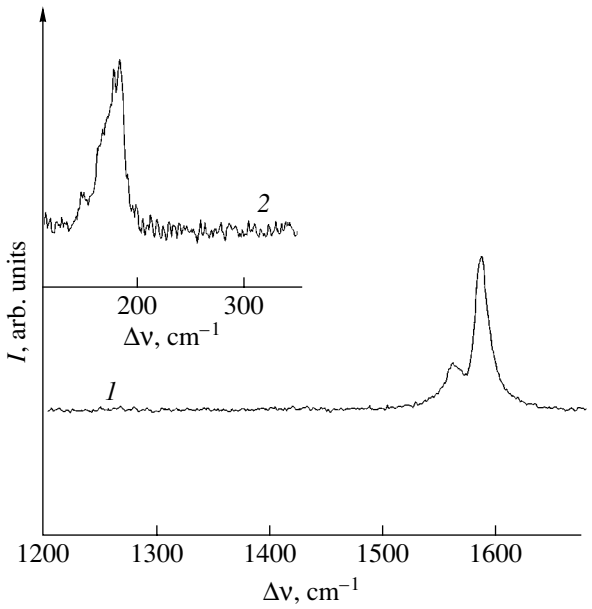


Fig. 5. Typical Raman spectra taken from the deposit with SWCNTs: (*1*) tangential mode and (*2*) radial mode.

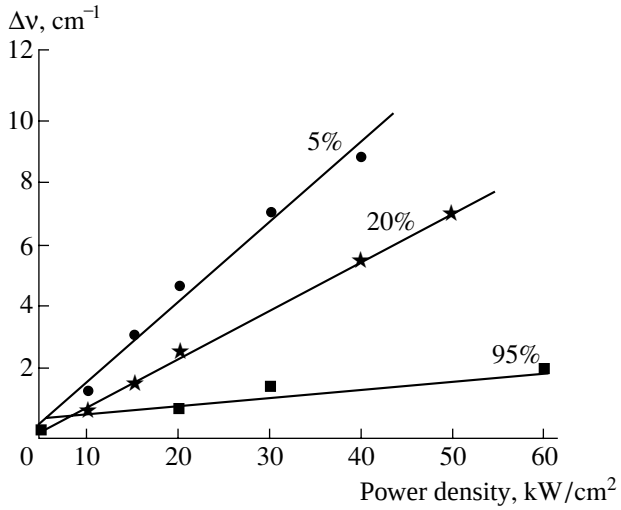


Fig. 6. Partial shift $\Delta\nu$ of the tangential mode peak vs. intensity of probing laser intensity (i.e., temperature) for a laser power of (●) 600 and (★) 800 W. (■) Sample prepared of purified SWCNTs [6].

in the stagnant zones of the reaction tube, where the flow velocity is low. Figure 8 demonstrates the micrograph of such a structure formed on the stainless steel grid. It appears that the filaments, which were 3 to 4 cm long in a number of experiments, consist of a large number of bundles of single-walled tubes.

MECHANISM OF SINGLE-WALLED CARBON NANOTUBE SYNTHESIS

The elucidation of an SWCNT formation mechanism and the elaboration of an adequate model of the

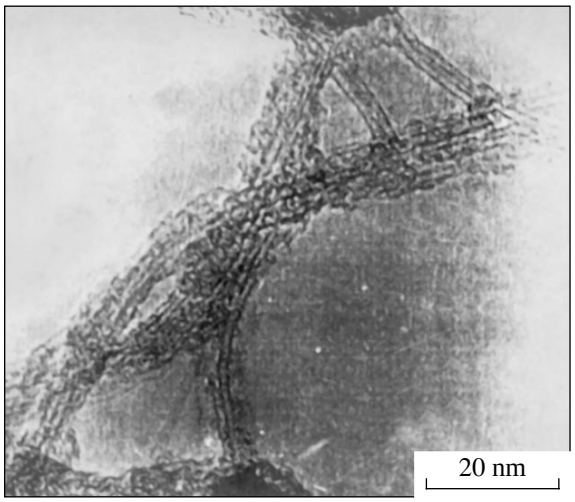


Fig. 7. HRTEM image of SWCNTs.

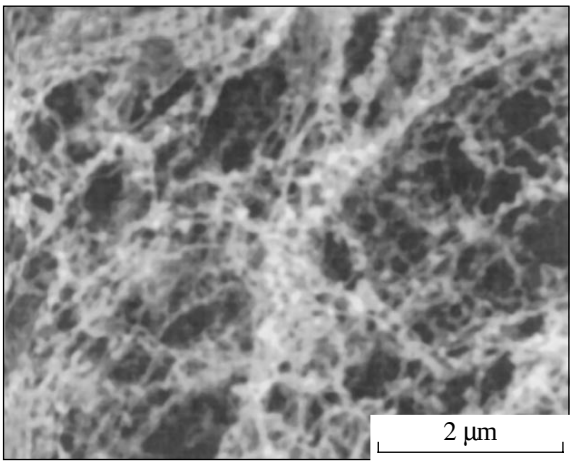


Fig. 8. SEM image of carbon webs deposited on the stainless steel grid.

process at high temperatures in a dense vapor–gas plasma environment are of great interest both for fundamental science and for the development of efficient growth technologies of nanostructures. We believe that, upon the laser ablation of the graphite target in the presence of the catalyst, SWCNTs nucleate and grow directly in the vapor–gas phase. The following experimental findings form the basis for the elucidation of the SWCNT formation mechanism in a laser torch.

(1) Note first of all that high-power-density laser radiation acting on condensed media causes intense evaporation with the characteristic expansion of ablation products and the formation of a vapor–gas plasma jet. In this jet (torch), electrons are readily produced primarily through the ionization of graphite, which has a low ionization potential (3.8 eV). Initially, the jet contains ions and atoms of all elements entering into the target composition, as well as molecular radicals and associates. In the peripheral (colder) parts of the jet, the radicals and associates are polymerized forming compact clusters. This is confirmed by mass spectrometry studies of the molecular composition of the saturated vapor of carbon, which revealed a set of molecules from C_2 to C_{10} [11]. In subsequent experiments on laser evaporation of carbon [12, 13], an even higher degree of carbon polymerization was observed. It was found that C_n macromolecules (n is an even number) are linear if $n < 8$, are cyclic if $12 < n < 28$, and form spherically closed structures (fullerenes) with $n > 50$. It should be noted that polymerization also takes place upon the laser evaporation of metal oxides. When yttria was exposed to pulsed laser radiation [14], $(Y_2O_3)_n(YO)^+$ ions, where n varies from 0 to 18, were detected.

(2) Graphite may easily dissociate with the formation of gaslike clusters $(–C\equiv C–)_n$, since a single bond in the graphene sheet breaks at high temperatures and the electron shifts toward a double bond, thereby inducing the break of the neighboring single bond. This, in turn, produces a triple bond in place of the former double bond. Similarly, an acetylene molecule with triple bonds is more stable against high temperatures.

(3) Next, it should be noted that any particle that is in contact with the plasma in the vapor–gas torch acquires a charge. Molecular associates and compact clusters may be viewed as large molecules, and their charge, which is apparently defined by the processes of electron attachment, ionization, and recombination, is no more than several elementary charges. A charged particle generates an electric field and induces the electric dipole moment in surrounding particles. If dipole–dipole interaction between particles is strong and the temperature is not too high, the effect of directed segregation may occur, causing interacting particles to align with the field and produce ordered linear structures.

Thus, the formation of a single-walled carbon nanostructure in our experiments may be considered as follows. When acting on a condensed medium, the intense laser radiation generates a vapor–gas plasma jet, which contains atoms, radicals, and carbon and catalyst molecules. At the edges of the jet, where the temperature is lower, the vapor condenses, forming molecular associates C_n and ionizing gaslike clusters that contain nickel, carbon, and yttrium. The resulting ions aggregate upon collisions, generating charged particles of a porous polymer cluster. The porous structure of an incipient cluster is seen in Fig. 7. From this figure it also follows that the porous structure of the cluster possibly with dangling bonds at the edges of the pores provides the formation of SWCNT bundles, which originate at the cluster surface. It seems that the diameter of nanotubes depends on the characteristic size of the porous structure.

In an alternative process, compact clusters forming at the periphery of the vapor–gas jet consist of carbon and catalyst atoms and molecules. According to a current point of view, Y_2O_3, when added to the catalyst,

prevents the clusters from increasing in size; therefore, they remain small and the resulting single-walled nanotubes have a small diameter. Since the temperature at the edges of the vapor–gas jet is lower, these clusters contain the supersaturated solution of carbon in nickel. As a result, excess carbon is precipitated on the cluster surface and serves as a seed for SWCNT growth.

The nanotubes grow in length through the interaction of a charged cluster nucleus with molecular associates provided that a sufficient quantity of building blocks is available in the environment. C_2 radicals, their associates, and chain structures mentioned above all may serve as building blocks being stable at high temperatures. In the field of a nucleus, molecular associates take an induced electric dipole moment and interact with the nucleus, producing strong chemical bonds due to the overlap of the carbon atom valence orbitals. Such bonding is typical of the graphene structure.

Also, a certain density of structure elements may trigger the mechanism of microscopic percolation due to dipole–dipole interaction between molecular associates. This mechanism produces an indefinitely long (critical) cluster (nanotube in our case). The occurrence of the mechanism is favored by sp^2 hybridized orbitals of carbon atoms located on the end face of a growing nanotube, which are extended normally to the nanotube axis. It is also obvious that the growth of SWCNTs may be interrupted at any instant because of the fluctuation formation of a defect (such as a pentagon with the subsequent formation of a fullerene-like "cap") in the crystal structure of the nanotube being synthesized.

To conclude, our results convincingly indicate that a laser power density of 5×10^4 W/cm^2 is optimal for the synthesis of single-walled carbon nanotubes under atmospheric pressure. The efficiency of the process naturally grows with increasing radiation power, and the fact that the process may go under atmospheric pressure greatly facilitates the synthesis. Note that various transition metals introduced into graphite (Fe, Co, Ni, etc.) perhaps should be named inductors, rather than catalysts (in view of their effect on the process), since they induce the synthesis in a definite direction.

A further series of experiments is aimed at optimizing the synthesis process and improving the quality of single-walled carbon nanotubes. Our goal is to use in full measure the advantage of the SWCNT laser synthesis, that is, the feasibility of the directed synthesis of carbon nanotubes with desired structure parameters.

ACKNOWLEDGMENTS

The authors thank N.A. Kiselev, E.D. Obraztsova, S.V. Terekhov, N.G. Solov'ev, and V.A. Kuznetsov for the valuable discussion and assistance.

This work was supported by the Russian Foundation for Basic Research (grant nos. 01-03-33025 and 02-01-00452) and the INTAS (grant no. 97-30810).

REFERENCES

1. S. Iijima, Nature **354**, 56 (1991).
2. W. K. Maser, E. Munoz, A. M. Benito, *et al.*, Chem. Phys. Lett. **292**, 587 (1998).
3. A. V. Eletskiĭ, Usp. Fiz. Nauk **172**, 401 (2002).
4. É. G. Rakov, Usp. Khim. **69** (1), 41 (2000).
5. G. I. Kozlov, Kvantovaya Élektron. (Moscow) **29**, 177 (1999).
6. E. D. Obraztsova, A. S. Pozharov, S. V. Terekhov, *et al.*, Mater. Res. Soc. Symp. Proc. **706**, 6 (2000).
7. É. I. Asinovskiĭ, A. V. Kirillin, and A. V. Kostanovskiĭ, Teplofiz. Vys. Temp. **35**, 716 (1997).
8. G. I. Kozlov, Zh. Tekh. Fiz. **72** (4), 81 (2002) [Tech. Phys. **47**, 453 (2002)].
9. S. Bandow, S. Asaka, Y. Saito, *et al.*, Phys. Rev. Lett. **80**, 3779 (1998).
10. A. Kasuya, Y. Sasaki, Y. Saito, *et al.*, Phys. Rev. Lett. **78**, 4434 (1997).
11. L. N. Sidorov, N. V. Korobov, and L. V. Zhuravlev, *Mass-Spectrometry Thermodynamic Studies* (Mosk. Gos. Univ., Moscow, 1985).
12. E. A. Rohlfing, D. M. Cox, and A. Caldor, J. Chem. Phys. **81**, 3322 (1984).
13. H. W. Kroto, J. R. Health, S. C. O'Brien, *et al.*, Nature **318**, 162 (1988).
14. F. Aubriet, N. Chaoui, R. Chety, *et al.*, Appl. Surf. Sci. **186**, 282 (2002).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1442–1448. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 83–89.

SOLID-STATE ELECTRONICS

Influence of Water Vapor Adsorption on the *C–V* Characteristics of Heterostructures Containing Porous Silicon

E. A. Tutov, E. N. Bormontov, V. M. Kashkarov, M. N. Pavlenko, and É. P. Domashevskaya

Voronezh State University, Universitetskaya pl. 1, Voronezh, 394693 Russia
e-mail: phssd2@main.vsu.ru

Received February 11, 2003

Abstract—Porous silicon (por-Si) is prepared by the electrochemical etching of single-crystal *n*-silicon in an aqueous–alcoholic solution of hydrofluoric acid in the presence of hydrogen peroxide oxidizer. The dependence of the high-frequency *C–V* characteristics of Al/por-Si/Si heterostructures on the relative humidity is studied. A model of capacitor structure is proposed, and a method of analyzing its capacitance as a function of the water vapor partial pressure in terms of the adsorption isotherm is elaborated. Within the framework of this model, the porosity of the material, the effective fraction of silicon dioxide in the por-Si, the fraction of intercommunicating porosity, the micropore-to-mesopore volume ratio, and the mesopore size distribution are determined. The porous silicon prepared in this work seems promising as a sensitive layer in capacitance-type humidity sensors. © *2003 MAIK "Nauka/Interperiodica"*.

INTRODUCTION

Integrated microelectronic sensors represent an advanced component basis for measuring, information, and control systems. Growing interest in this area of research is due to the fact that the efficiency of existing computerized systems is limited by the performance of data acquisition and processing devices rather than by the hardware and software capabilities.

Humidity sensors occupy a prominent place among pickups of various types. The measurement of the humidity of gas mixtures, for example in air, is of great importance for scientific, industrial, and medical application [1].

High-end silicon technology makes it possible to produce integrated sensors. Silicon cannot be considered as the best material for sensitive elements; however, it allows for integration with other electronic elements on a single chip, which offers considerable advantages over the traditional approaches.

Porous silicon has attracted considerable attention mainly owing to its luminescent properties [2]. The subsequent extensive study of this material has demonstrated its promise for biotechnology, solar cells, and sensors [3].

The adsorption properties of porous silicon may be optimized by electrochemical etching. The effects of adsorption of various gases, including water vapor, on the electrical characteristics of por-Si and the use of these effects in resistive-, diode-, and capacitance-type sensors were investigated in [4–8].

In this work, we study the response of a capacitance-type sensor to a change in the relative humidity and also consider the inverse problem, namely, the determination of structural and phase parameters of por-Si (the oxide phase fraction, the porosity, and the pore size distribution) from the variation of the capacitance of an Al/por-Si/*c*-Si heterostructure with water vapor pressure.

ADSORPTION CAPACITANCE POROSIMETRY

The features of real porous systems as objects for structural–phase analysis, the present concepts of their structure, and the progress in the methods of their production allow one to consider porosimetry as a separate area of structural analysis. Nowadays, about sixty methods for studying the porous structure of solids are known (pycnometry, mercury porosimetry, adsorption structural analysis, microscopic methods, calorimetry, etc.) [9]. They provide much information, offer a high sensitivity, and have a wide range of application [10]. All of them handle macroscopic amounts of analytes and provide information on the integral porosity. For studying highly porous materials with a considerable fraction of micropores, adsorption structural analysis, which uses the adsorption isotherm as the initial source of information, is the method of choice.

In recent years, much attention has been given to porous films of anodic oxides of metals (Al_2O_3 [11]) and semiconductors (e.g., por-Si [3]), which are of considerable scientific and practical interest. The nature of these films does not allow the wide use of volumetric or gravimetric adsorption methods; however, the electrical parameters of heterostructures with a porous anodic layer of a metal or semiconductor are highly sensitive

1063-7842/03/4811-1442$24.00

to external factors, in particular, to the adsorption of water vapor.

The choice of water vapor was dictated by the anomalously high permittivity of water ($\varepsilon = 81$) rather than by a small kinetic diameter ($d = 0.264$ nm [9]) and small water molecule adsorption area ($S = 0.125$ nm^2 [12]). The first circumstance is bound to change significantly the effective permittivity of the porous layer after water vapor adsorption.

A *c*-Si/por-Si heterostructure forms during the growth of the porous layer in a natural way. Therefore, to make a capacitor only requires the deposition of a metallic contact of a given size. At the same time, a large dipole moment of a water molecule increases the energy of water–por-Si interaction and allows one to neglect the adsorption of N_2, CO_2, and O_2 molecules from the air.

The measurement of the capacitance of a metal–anodic layer–metal structure (combined with the measurement of the high-frequency *C–V* characteristic for a semiconductor–anodic layer–metal structure) as a function of the water vapor pressure not only enables one to determine the properties of the porous anode film but also provides for the otherwise impossible locality of analysis.

The dependence of semiconductor properties on the atmospheric composition was noted in a number of studies. Those investigations were aimed mainly at analyzing the modification of the properties due to the material–environment interaction. The inverse idea is to determine the composition of the gaseous atmosphere around semiconductors through changes in their electrical characteristics [13]. Our technique also tackles the third aspect of the problem: studying the properties of a semiconductor adsorbent itself by analyzing its response to a change in the ambient air. Let us discuss this technique, which may be given the name adsorption capacitance porosimetry, in detail.

According to the model representation of the porous silicon structure formation [14] and to the experimental data on the morphology of por-Si layers [15], the porous layer may be considered (under certain anodizing conditions and with insignificant simplifications) as an irregular network of cylindrical pores that have the same diameter but differ in length. The pores pierce the silicon matrix. Note that, for cylindrical pores, the definitions of surface and volume porosity coincide.

It is known that porous silicon obtained by anodic oxidation contains unoxidized silicon and silicon in the oxidation states Si^{2+} and Si^{4+}, i.e., the SiO and SiO_2 compounds [16, 17], with the oxide fraction decreasing from the surface to the bulk of the layer. The deposition of a metallic electrode (for example, aluminum) on the porous silicon surface forms an Al/por-SiO_x/por-Si/*c*-Si capacitor heterostructure.

At a relative humidity RH = 0, the capacitance of this structure in the enhancement mode, C^0_{max}, asymptotically tends to the geometrical capacitance of a porous insulating layer with an effective permittivity ε_{eff} and has two components: the capacitance of the matrix with a permittivity ε_{SiO_x} and that of air-filled pores with a permittivity ε_{air}. Therefore, under the assumption that $\varepsilon_{SiO_x} = \varepsilon_{SiO_2}$, the effective permittivity of a porous insulating layer is given by

$$\varepsilon_{eff} = \varepsilon_{SiO_2}(1 - P) + \varepsilon_{air}P, \quad (1)$$

where $P = S_{por}/S$ is the porosity ratio, S is the gate dielectric surface area, and S_{por} is the surface area of pores on the por-Si surface under the electrode. The geometrical capacitance C^0_{max} of the capacitor depends on the effective thickness d_{eff} and effective permittivity ε_{eff} of the gate dielectric:

$$C^0_{max} = \frac{\varepsilon_{eff}S}{d_{eff}}. \quad (2)$$

In the depletion mode, the total high-frequency (HF) capacitance of the structure decreases, since the capacitance of the space charge region in the silicon connects in series to the geometrical capacitance of the insulating layer, and achieves a minimum C_{min} under strong inversion:

$$\frac{1}{C_{min}} = \frac{1}{C^0_{max}} + \frac{1}{C_{sc_{min}}}. \quad (3)$$

Here,

$$C_{sc_{min}} = \frac{\varepsilon_{Si}(1 - P)S}{W_m} \quad (4)$$

is the minimal HF capacitance of the space charge region (SCR) in the silicon and W_m is the SCR maximal width at a given impurity concentration N [18]:

$$W_m = \sqrt{\frac{4\varepsilon_{Si}kT\ln(N/n_i)}{q^2N}}, \quad (5)$$

where q is the electron charge, k is the Boltzmann constant, T is absolute temperature, ε_{Si} is the absolute permittivity of silicon, and n_i is the intrinsic carrier concentration.

Using the experimental values of C^0_{max} and C_{min} and of the impurity concentration N in the silicon, we find from equations (3)–(5) the porosity ratio P. Substituting the obtained value of P into Eq. (1) yields the effective permittivity ε_{eff} of the insulating layer. Finally, substituting ε_{eff} into Eq. (2), we find the effective thickness d_{eff} of the insulator.

In the presence of water vapor, the total capacitance C_{max} of the enhancement-mode heterostructure can be represented as the sum of three terms: the capacitance of the matrix with the permittivity ε_{SiO_2}, the capaci-

tance of air-filled pores with the permittivity ε_{air}, and that of condensate-filled pores with the permittivity ε_{H_2O} (in view of the partial penetration of the water vapor under the metallic electrode):

$$C_{max} = \frac{S}{d_{eff}}[\varepsilon_{SiO_2}(1-P) + \varepsilon_{air}(Pk(1-\eta) + P(1-k)) + \varepsilon_{H_2O}Pk\eta]. \quad (6)$$

Here, $\eta = S_{fill}/S_{por}$ is the fill of pores with the condensate, S_{fill} is the surface area of condensate-filled pores on the por-Si surface under the electrode at a given vapor pressure, and k is the ratio between the surface area of pores accessible for the water vapor and the total surface area of pores under the electrode (the pore accessibility factor).

The capacitance of the structure at 100% humidity increases because of the capillary condensation of the vapor in micro- and mesopores at the aluminum electrode boundary. At RH = 100% ($\eta = 1$), expression (6) becomes

$$C_{max}^{100} = \frac{S}{d_{eff}}[\varepsilon_{SiO_2}(1-P) + \varepsilon_{air}P + (\varepsilon_{H_2O} - \varepsilon_{air})Pk] = \frac{S}{d_{eff}}[\varepsilon_{eff} + (\varepsilon_{H_2O} - \varepsilon_{air})Pk]. \quad (7)$$

A relation between the maximal capacitance of the structure at RH = 0 (C_{max}^0) and that at RH = 100% (C_{max}^{100}),

$$C_{max}^{100} = C_{max}^0\left[1 + \frac{\varepsilon_{H_2O} - \varepsilon_{air}}{\varepsilon_{eff}}Pk\right], \quad (8)$$

enables one to determine the coefficient k ($0 < k < 1$), which depends on the surface area-to-perimeter ratio of the electrode, total porosity, fraction of intercommunicating porosity, and, possibly, on the relative humidity. This coefficient is the upper estimate of the intercommunicating porosity, since adsorption in a system of interconnected pores results in the filling of all pores under the electrode; i.e., $k = 1$ whatever the capacitor geometry.

The fill of pores with a condensate, η, can be determined from Eq. (6), which, in view of Eqs. (8) and (2), takes the form

$$C_{max} = C_{max}^0 + (C_{max}^{100} - C_{max}^0)\eta = C_{max}^0 + \frac{(\varepsilon_{H_2O} - \varepsilon_{air})S}{d_{eff}}Pk\eta. \quad (9)$$

Thus, from experimental data for the high-frequency capacitance of an enhancement-mode Al/por-Si/*c*-Si structure with RH = 0, 100%, and 0 < RH < 100% (C_{max}^0, C_{max}^{100}, and C_{max}, respectively), as well as of the inversion-mode structure (C_{min}), one can calculate the total porosity P, the fraction k of intercommunicating porosity, and the effective parameters ε_{eff} and d_{eff} of the oxide phase in the porous layer.

Now let us turn to the analysis of the pore size distribution. Depending on the nature and production of solid adsorbents, they may contain pores of different size and morphology. The transverse size of the pores is of special interest in most cases. The classification of pores by size that is currently adopted by the International Union of Pure and Applied Chemistry (IUPAC) assigns each range of pore sizes to a specific mechanism of adsorption [19]. According to this classification, porous silicon may be microporous (a pore diameter $D < 2$ nm, the mono- and polymolecular adsorption mechanisms), mesoporous ($2 < D < 50$ nm, capillary condensation), and macroporous ($D > 50$ nm, the adsorption properties are different from those of the nonporous material only slightly). Today's technology allows the formation of por-Si layers with different pore size distribution; therefore, the pickups may be made sensitive to different ranges of humidity [6].

The amount of water adsorbed in micropores can be estimated with the BET equation [19]

$$\frac{n}{n_m} = \frac{1}{1 - p/p_0} - \frac{1}{1 + (p/p_0)(Q-1)}. \quad (10)$$

Here, n is the absolute amount of an adsorbate (in moles), n_m is the amount of the adsorbate per monolayer (in moles), $Q \approx \exp[(q - q_L)/RT]$ (where q is the heat of adsorption and q_c is the heat of condensation; i.e., $q - q_c$ is the pure heat of adsorption), R is the gas constant, T is the temperature, p is the pressure, and p_0 is the saturation vapor pressure. Equation (10) describes polymolecular adsorption and defines the number of layers adsorbed at a given pressure.

The adsorption mechanism typical of mesopores is capillary condensation, which is described by the Kelvin equation [19]

$$p = p_0\exp\left(\frac{2\sigma V_m}{rRT}\right), \quad (11)$$

where p_0 is the saturation vapor pressure at a temperature T, σ is the surface tension coefficient of the liquid phase (water), V_m is the molar volume of the liquid phase, and r is the capillary radius (negative in the case of a wetting liquid).

This equation implies that, for a given relative pressure p/p_0, there is a critical radius r_{cr} such that all pores with a radius smaller than r_{cr} are filled.

One may expect that the pore radius r correlates with the porous layer formation conditions. In simulating the pore size distribution $f(r)$, one most often uses

the logarithmically normal distribution [19]

$$f(r) = \frac{1}{\sigma_g\sqrt{2\pi}}\exp\left[-\frac{(\ln r - \ln r_g)^2}{2\ln^2\sigma_g}\right], \quad (12)$$

where r_g is the geometric mean of r and $\ln\sigma_g$ is the standard deviation of $\ln r$ (the spread in the values about the mean).

Let us find the pore size distribution within the framework of the capacitance porosimetry method (discussed in this study) from the experimental dependence of capacitance C_{max} (6) on the relative humidity p/p_0. Note that the relative humidity p/p_0 enters (6) only through the parameter η defined as

$$\eta = \frac{S_{fill}}{S_{por}} = \frac{1}{S_{por}}\int_0^{r_k} f(r)\pi r^2 dr. \quad (13)$$

Here, $r_{cr} = r_{cr}(p/p_0)$ is the critical pore radius, which is calculated from the Kelvin equation and specifies the maximum size of condensate-filled pores at a given humidity, and $f(r)$ is the pore size distribution.

Differentiating dependence $C_{max}(p/p_0)$ (6) as a complex function in view of Eqs. (11) and (13), we have

$$\frac{dC_{max}}{d(p/p_0)} = \frac{dC_{max}}{d\eta}\frac{d\eta}{dr}\frac{dr}{d(p/p_0)} = \frac{\varepsilon_{H_2O} - \varepsilon_{air}}{d_{eff}}\frac{k\pi r^4 RT}{2\sigma V_m}\exp\left(\frac{2\sigma V_m}{rRT}\right)f(r). \quad (14)$$

From (14), one finds the exponential mesopore ($r >$ 1.2 nm) size distribution

$$f^*(r) = \frac{d_{eff}}{\pi k(\varepsilon_{H_2O} - \varepsilon_{air})}\frac{2\sigma V_m}{RT}\frac{1}{r^4}\exp\left(-\frac{2\sigma V_m}{rRT}\right)\frac{dC_{max}}{d(p/p_0)} \quad (15)$$

by graphically differentiating the experimental dependence $C_{max}(p/p_0)$.

The remaining part of the function $f(r)$ (for micropores) is found with model distribution (12) normalized to the total porosity and BET equation (10).

Thus, the physical adsorption of water vapor in micropores and its capillary condensation in mesopores increase the permittivity of the porous layer, which can be utilized in capacitance-type humidity sensors. The analysis of the capacitance of such a sensor vs. relative humidity in terms of the adsorption isotherm makes it possible to find the ratio between the concentrations of micro- and macropores in silicon, as well as the size distribution of the latter.

RESULTS AND DISCUSSION

The objects of investigation were porous silicon samples whose electrical characteristics responded to the relative humidity variation in the range from 0 to 100%.

To form the porous layer, a phosphorus-doped single-crystal (100)Si wafer with an electron concentration of 1.3×10^{16} cm^{-3} was rinsed in distilled water and then cleaned in an HF + HNO_3 + CH_3COOH etchant. Anodizing was carried out in an HF/H_2O : C_3H_8O : H_2O_2 = 2 : 2 : 1 electrolyte at a current density of 15 mA/cm^2 for 5 min. The samples anodized were rinsed in butyl alcohol and dried. The surface of the anodized wafers remained mirror, thus indicating the absence of micropores. Finally, 2-mm^2 aluminum contacts were thermally evaporated.

An electron micrograph from the transverse cleavage surface of the structure is shown in Fig. 1a. The por-Si layer has a uniform columnar structure and a thickness of about 10 μm, and the transition layer at the por-Si/*c*-Si interface is about 10% of the total thickness.

Stationary HF *C*–*V* characteristics were taken with a computerized setup [20] (the frequency and amplitude of a test signal were 1 MHz and 20 mV). Each curve was taken for about 5 min. The leakage current was estimated as no more than 1 mA. The wiring capaci-

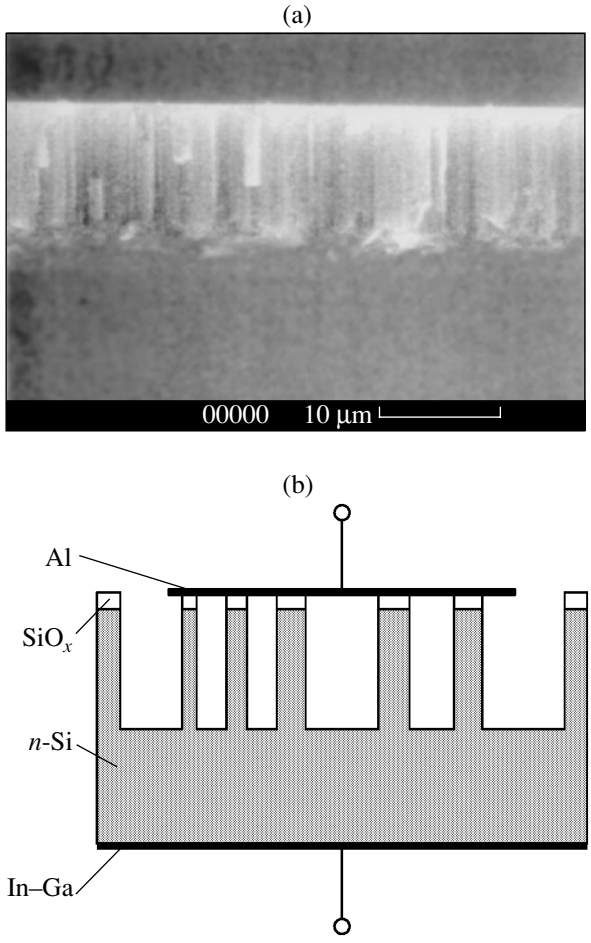


Fig. 1. (a) Electron micrograph of the transverse cleavage surface and (b) schematic of the Al/por-Si/*c*-Si heterostructure.

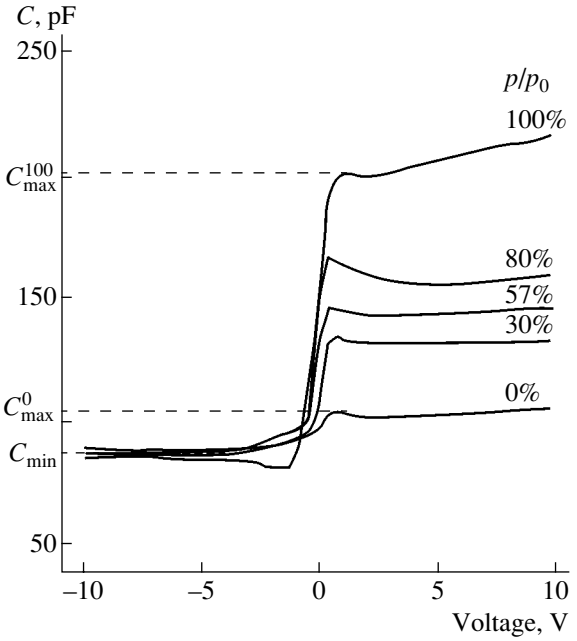


Fig. 2. High-frequency *C–V* characteristic of the Al/por-Si/*n*-Si structure with the relative humidity varying in the range 0–100%.

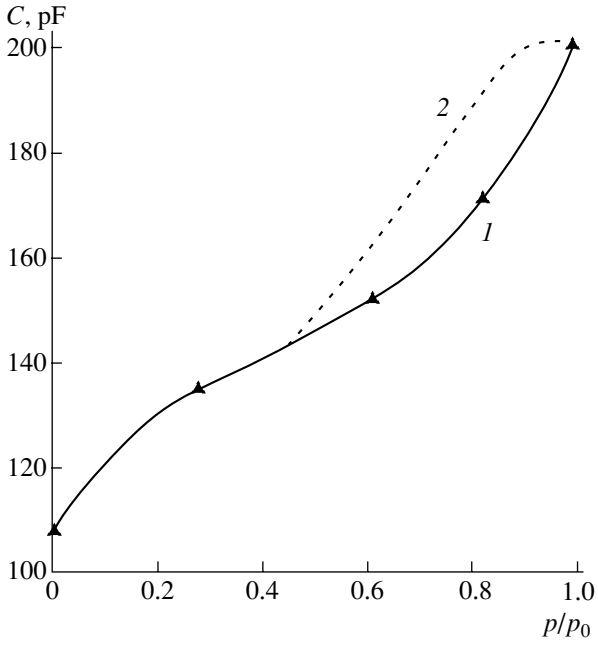


Fig. 3. Capacitance of the sensor vs. the relative humidity. (▲) Experiment and (*1*, *2*) calculation for the exponential (see (15)) and model (see (12)) pore size distribution.

tance (≈1 pF) was automatically subtracted from the measurements. The test sample was placed into a pressure-tight measuring cell, which was dried out with ShSM silica gel, and heated for 30 min at $T = 343$ K. The residual water vapor pressure was assumed to correspond to zero relative humidity (RH = 0). The vapor saturation regime (RH = 100%) was established in equilibrium with the open distilled water surface.

The RH values (30, 57, 80, and 100%) were set with an aqueous solution of glycerol taken in an appropriate proportion. This technique is simpler and more flexible than the techniques using the saturated solutions of salts [21]. The amount of the solution was chosen such that one may neglect the variation of its concentration during gas exchange with the measuring cell volume.

The HF *C–V* characteristics of the Al/por-Si/*n*-Si structure for different RH values are shown in Fig. 2. Their shape is typical of MIS structures with a low density of surface states and a small amount of charge built in the insulator. The physical adsorption of water may affect neither the built-in charge nor the effective concentration of free carriers in the silicon matrix. Indeed, at a negative bias, the capacitance changes (increases) only slightly and the *C–V* curve does not shift along the voltage axis. At a positive bias, the capacitance grows with RH because of an increase in the effective permittivity of the porous layer due to the water vapor adsorption. The slight deviation of the *C–V* curve from the general run at RH = 100% is related to the fact that this curve is essentially nonequilibrium because of a higher conductivity of the structure in this case.

Using the experimental values of the enhancement-mode capacitances C^0_{max} and C^{100}_{max}, the value C_{min} measured under inversion (Fig. 2), and formulas (1)–(7), one can calculate the total porosity ($P = 0.5$) and effective parameters of the oxide phase in the porous layer: the thickness ($d_{eff} = 0.4$ μm) and relative permittivity ($\varepsilon_{eff} = 2.4$). Then, the fraction k of intercommunicating pores is found from (8). Our calculation yields $k = 0.06$; i.e., most pores in the por-Si do not communicate. The adsorption capacitance porosimetry data show the

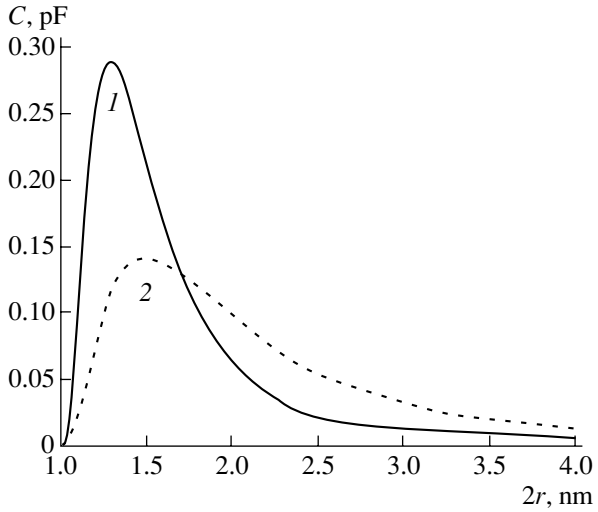


Fig. 4. Pore size distribution: (*1*) calculation by formula (15) and (*2*) logarithmically normal model distribution.

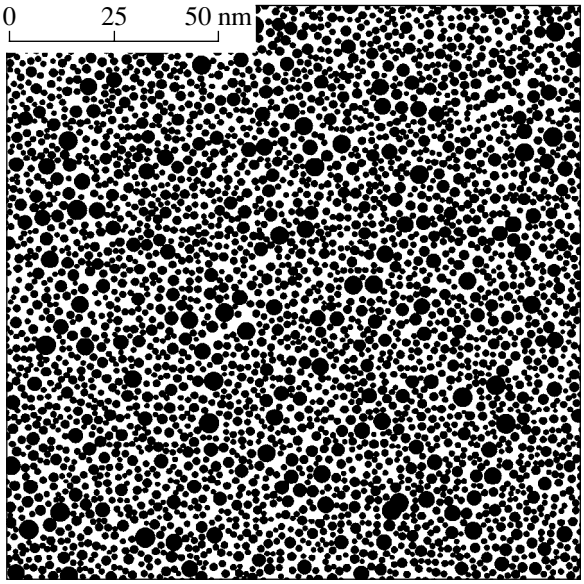


Fig. 5. Simulated porous silicon surface for a porosity of 50% and pore size distribution shown by curve *1* in Fig. 4. Pores are shown black.

spread in the structural–phase parameters of the porous silicon over the wafer surface: the porosity ranges from 50 to 70% and the effective thickness of the oxide phase (SiO_2) varies between 0.3 and 0.4 μm (i.e., within 3–4% for a 10 μm-thick porous layer).

The porous layer capacitance in the positively biased structure (enhancement mode) versus the relative humidity is shown in Fig. 3. Calculation of the pore size distribution from the experimental dependence $C(p/p_0)$ is possible for only capillary condensation in mesopores, since the Kelvin equation does not describe adsorption in micropores. Therefore, only when the volume (and capacitance) of micropores is subtracted from the total volume of pores can the mesopore size distribution and the fraction of micropores in the total porosity be determined. For water, the minimal mesopore radius equals 1.2 nm, which corresponds to condensation at a relative humidity of 35% [19]. The dashed curve in Fig. 3 shows the capacitance–humidity dependence calculated with regard for the circumstance mentioned above.

The general form of the pore size distribution found from (12) and (15) and taking into account the fraction of micropores is presented in Fig. 4. Under our anodizing conditions, the diameter of most pores falls into a narrow range between 1 and 4 nm and the fraction of micropores accounts for 40% of the total volume of pores. Macropores, if any, give a minor contribution to the statistics of water vapor adsorption and desorption; however, their transporting role remains crucial for the kinetic parameters of humidity sensors [6]. The appearance of the simulated por-Si surface with the pore size distribution obtained for $P = 0.5$ is depicted in Fig. 5.

Note that investigation into the kinetic characteristics of capacitance-type sensors goes beyond the scope of this study. However, it should be mentioned that physical adsorption is a fast process and capillary condensation is characterized by hysteresis. The optical properties of porous silicon suggest that radiations with different wavelengths may be used to selectively excite por-Si electron subsystems and, thus, control the response and relaxation of structure [22]. Thus, many properties of porous silicon can be optimized, which makes it a candidate for microelectronic sensors.

CONCLUSIONS

The electrochemical etching of *n*-type single-crystal (100) silicon in an aqueous–alcohol solution of hydrofluoric acid in the presence of an oxidizer produces a porous silicon layer where micro- and mesopores occupy 70% of the volume. Three to four percent of the volume is accounted for by the oxide phase. The physical adsorption of water vapor in micropores and the capillary condensation in mesopores increase the permittivity of the layer, which opens a way to the production of capacitance-type humidity sensors that respond to RH = 0–100%. Analyzing the por-Si capacitance as a function of the relative humidity in terms of the adsorption isotherm, one gains information on the phase–structural characteristics of porous silicon, even if the layer has a more complex morphology.

REFERENCES

1. B. I. Podlepetskiĭ and A. V. Simakov, Zarubezhn. Élektron. Tekh., No. 2, 64 (1987).
2. K. H. Jung, S. Shih, and D. L. Kwong, J. Electrochem. Soc. **140**, 3046 (1993).
3. *EMIS Data Review Ser.* no. 18, *Properties of Porous Silicon*, Ed. by L. Canham (IEE, London, 1997).
4. V. M. Demidovich, G. B. Demidovich, E. I. Dobrenkova, and S. N. Kozlov, Pis'ma Zh. Tekh. Fiz. **18** (14), 57 (1992) [Sov. Tech. Phys. Lett. **18**, 459 (1992)].
5. V. M. Demidovich, G. B. Demidovich, S. N. Kozlov, *et al.*, Vestn. Mosk. Univ., Ser. 3: Fiz., Astron., No. 4, 99 (1996).
6. A. A. Kovalevskiĭ and I. L. Baranov, Mikroélektronika **25**, 298 (1996).
7. R. C. Anderson, R. S. Muller, and C. W. Tobias, Sens. Actuators A **23**, 835 (1990).
8. T. Taliercio, M. Dilhan, E. Massau, *et al.*, Sens. Actuators A **46** (1–3), 43 (1995).
9. T. G. Plachenov and S. D. Kolosentsev, *Porosimetry* (Khimiya, Leningrad, 1988).
10. P. G. Cheremskoĭ, *Methods of Studying the Porosity in Solids* (Énergoatomizdat, Moscow, 1985).
11. V. V. Petrova, *Microporosity of the Anode Oxide Aluminum Films* (Petrozavodsk. Univ., Petrozavodsk, 1992).
12. A. L. McClellan and H. E. Harnsberger, J. Colloid. Interface Sci. **23**, 577 (1967).

13. I. A. Myasnikov, V. Ya. Sukharev, L. Yu. Kupriyanov, and S. A. Zav'yalov, *Application of Semiconductor Sensors for Physicochemical Research* (Nauka, Moscow, 1991).
14. L. N. Aleksandrov and P. L. Novikov, Pis'ma Zh. Éksp. Teor. Fiz. **65**, 685 (1997) [JETP Lett. **65**, 714 (1997)].
15. É. Yu. Buchin, A. V. Postnikov, A. V. Prokaznikov, *et al.*, Pis'ma Zh. Tekh. Fiz. **21**, 60 (1995) [Tech. Phys. Lett. **21**, 27 (1995)].
16. E. P. Domashevskaya, V. M. Kashkarov, E. Yu. Manukovskii, *et al.*, J. Electron Spectrosc. Relat. Phenom. **88–91**, 958 (1998).
17. L. V. Belyakov, T. L. Makarova, and V. I. Sakharov, Fiz. Tekh. Poluprovodn. (St. Petersburg) **32**, 1122 (1998) [Semiconductors **32**, 1003 (1998)].
18. S. Sze, *Physics of Semiconductor Devices* (Wiley, New York, 1981; Mir, Moscow, 1984).
19. S. Gregg and K. Sink, *Adsorption, Surface Area and Porosity* (Academic, New York, 1982; Mir, Moscow, 1984).
20. E. N. Bormontov and S. V. Golovin, Izv. Vyssh. Uchebn. Zaved. Élektronika, No. 4, 95 (1998).
21. G. Wiegleb, *Sensortechnik* (Franzis, Munchen, 1986; Mir, Moscow, 1989).
22. E. A. Tutov, A. Yu. Andryukov, and S. V. Ryabtsev, Pis'ma Zh. Tekh. Fiz. **26** (17), 53 (2000) [Tech. Phys. Lett. **26**, 778 (2000)].

Translated by A. Sidorova

Technical Physics, Vol. 48, No. 11, 2003, pp. 1449–1453. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 90–94.

SOLID-STATE ELECTRONICS

Electronic Structure of Neutral Silicon–Vacancy Complex in Diamond

S. S. Moliver

Ul'yanovsk State University, Ul'yanovsk, 432700 Russia

e-mail: moliver@sv.uven.ru

Received April 24, 2003

Abstract—Quantum-chemical open-shell method with parametrization INDO is used to construct the model of a large unit cell for the $[Si–V]^0$ impurity complex in diamond. It is shown that the ground spin-triplet state of the complex exerts tunnel (rather than Jahn–Teller) splitting associated with off-center shift of Si atom along the trigonal axis of the fully symmetric atomic configuration $\mathbf{D}_{3d}$. Therefore, this complex, being a source of electron spin resonance *KUL*1 $S = 1$, may appear to be the known optical 1.68-eV center with a typical (≈1 meV) splitting of the zero-phonon line. The intracenter optical excitation occurs from the filled orbital doublet localized at Si to the orbital doublet localized at three of six carbon atoms neighboring Si and has a multiplet structure. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The impurity silicon–vacancy (*V*) complex in diamond was first identified in [1]. Studying the fine structure of its optical absorption and luminescence (this center acquires the spectroscopic denotation of 1.68-eV center in reference to the line of 737-nm zero-phonon electron transition) revealed the splitting by values of 0.20 and 1.07 meV, which are typical of tunnel behavior or of the dynamic Jahn–Teller effect [2], although the spin–orbit interaction [3] cannot be excluded from possible causes of splitting. Available experimental data are yet insufficient to determine the symmetry of the center; the measurements of the excited-state lifetime yield only implicit information on the structure of the center's closest neighborhood in diamond films [4, 5].

The 1.68-eV center is responsible for one of the main peaks in the optical spectra of CVD diamond films. However, in spite of considerable theoretical and experimental effort, it is not clear whether it is also an "$S = 1$" center, where S is the total spin of localized electrons. With the use of the density functional theory [6], it is predicted that the neutral charge state of the complex $[Si–V]^0$ is spin-triplet; i.e., this is an $S = 1$ center. It is only recently that the optical absorption and luminescence measurements, along with the electron spin resonance (ESR) spectroscopy [7], allowed one to establish the two characteristic signals with the determined principal values of spin Hamiltonian tensor [8]: these are the *KUL*1 ($S = 1$, $\mathbf{C}_{3v}$) signal for the neutral state and *KUL*8 ($S = 1/2$) signal for the state with charge –1. Similar calculation [6] also suggests that the fully symmetric $\mathbf{D}_{3d}$ and orbitally nondegenerate $^3A_{2g}$ $[Si–V]^0$ center cannot be split by optical excitation; i.e., it is not the 1.68-eV center. In contrast, when the center is in the –1 state, the energy of $^2E_g \longrightarrow {}^2E_u$ transition calculated in [6] is close to 1.68 eV and the Jahn–Teller effect takes place; as a result, it was concluded that the $[Si–V]^-$ complex may appear to be the 1.68-eV center.

STATEMENT OF THE PROBLEM

In order to interpret the experimental properties of [Si–*V*], let us calculate the singlet–triplet splitting and the structural barriers in terms of quantum-chemical approach, that is, by the method of molecular orbitals (MOs) in a restricted open-shell Hartree–Fock (ROHF) approximation with intermediate neglect of differential overlap (INDO) semiempirical parametrization [9]. Formally, the calculation of the electron and atomic structure of this complex closely resembles that in the case of divacancy [10] (Fig. 1a).

(1) Fully symmetric $\mathbf{D}_{3d}$ state results from the location of an impurity Si atom at the saddle point of vacancy migration, so that the center may be called "Si half-vacancy"; the same is implied when this complex [8] is referred to as *V*Si*V*.

(2) In the presence of six broken bonds of carbon and four broken bonds of silicon, the calculation yields (the same result is obtained with other calculation techniques and defect models, e.g., [6]) an open electron shell e_g^2 with ROHF coefficients (see table). These coefficients are included in the self-consistent calculation of the total energy of terms or diagonal Slater sums E" by the open-shell method [10].

(3) The resulting possibilities for the ground state of $[Si–V]^0$ are as follows: (a) Jahn–Teller effect for zero-spin electron doublet reduces the symmetry of the center to $\mathbf{C}_{2h}$ with the splitting of doubly degenerate levels; (b) the exchange interaction between two electrons of

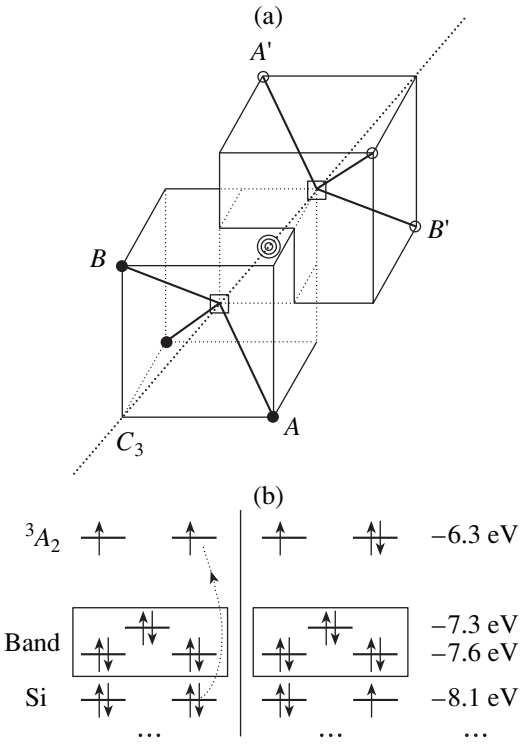


Fig. 1. (a) The structure of the complex: circles and squares stand for carbon atoms and carbon vacancies at crystal sites, Si atom is shown by concentric circle in the vicinity of the saddle point in the migration path from an empty site (semi-vacancy) to the nearest site (the other half-vacancy) along the selected [111] trigonal axis; (b) scheme of the optical intracenter excitation; this scheme semiquantitatively corresponds to the absorption and luminescence at 1.68 eV.

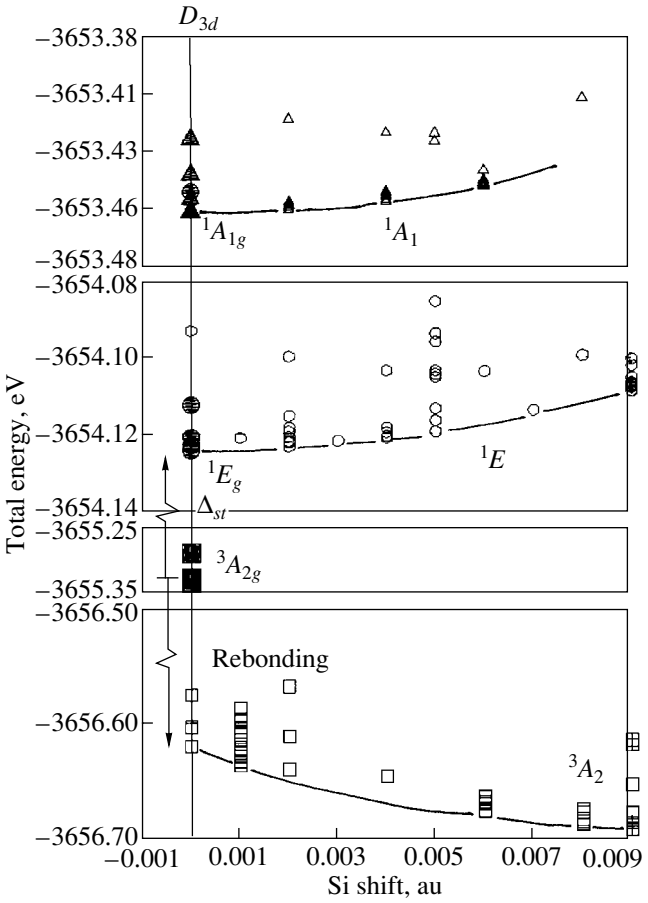


Fig. 2. Multiplet structure of the center. Abscissa shows the shift of Si atom along trigonal axis (arbitrary units times the factor 6.3 gives the value in angstroms). Different stages of energy minimization via the relaxation of carbon atom triads to Si and to half-vacancies are shown. The larger and darker symbols denote the energies of totally symmetric configuration.

the open shell is sufficiently strong to remove the orbital degeneracy, and the ground state is the totally symmetric spin triplet $^3A_{2g}$; and (c) Si atom shifts along the trigonal axis, the symmetry reduces to $\mathbf{C}_{3v}$, but the open-shell degeneracy is retained (off-center position [11]).

CALCULATION OF MULTIPLET STRUCTURE

The simulation is carried out for a 31-atom large unit cell (LUC). The main result of this calculation, the self-consistent energies of many-electron terms (see table), is shown in Fig. 2. The equilibrium relaxation of the six nearest neighbors of silicon is very small (0.04–0.05 Å) and depends on the term ($^1A_{1g}$, ^1E_g, $^3A_{2g}$), which can be attributed to a large covalent radius of impurity Si and the strength of the diamond crystal bonds. The open-shell molecular orbitals are entirely composed of six broken carbon bonds; i.e., Si impurity acts as a chemical acceptor when it transfers eight of total ten electrons of dangling bonds to the valence band. The corresponding resonance levels including no less than 10% (by density) of Si atomic orbitals (AO) are marked by triangles in the electron spectrum of the valence states (Fig. 3).

Figure 2 shows the dependence of the multiplet structure on the off-center shift of the impurity Si as it shifts along the trigonal axis towards the three nearest neighbors: in quantum chemistry terms, this is a resonant bonding with the reduction of symmetry $\mathbf{D}_{3d} \longrightarrow \mathbf{C}_{3v}$. Total-energy minimization was performed via the relaxation of these three neighboring carbon atoms and, then, the other three atoms from which the Si atom moves away, to "their" half-vacancy.

In fully symmetric atomic configuration, the lowest term is spin-triplet: $^3A_{2g} \approx {}^1E_g - 1.2$ eV. Like in the calculations for other vacancy-containing centers (vacancy, divacancy), the value of the spin-triplet splitting Δ_{st} obtained by the MO method is overstated because of a lower correlation energy of spin-triplet state as compared to that of a single state. In a center with the same symmetry of a half-vacancy [12], $\Delta_{st} \approx 1.5$ eV,

The multiplet structure and the coefficients of half-filled doubly degenerate open shell [10]

e_g^2, e_u^2 $\mathbf{D}_{3d}$	e^2 $\mathbf{C}_{3v}, \mathbf{D}_3$	e_g^2, e_u^2 $\mathbf{S}_6$		A_J	A_K	A_I
$^3A_{2g}$	$^3A_2$	^3A_g	$e'e''\alpha\alpha,\ e'e''\beta\beta,\ \frac{1}{\sqrt{2}}\,e'e''(\alpha\beta+\beta\alpha)$	1	2	0
		^1A_g	$\frac{1}{\sqrt{2}}\,e'e''(\alpha\beta-\beta\alpha)$	1	–2	–4
$^1A_{1g}$	$^1A_1$		$\frac{1}{\sqrt{2}}\,(e'e'+e''e'')\alpha\beta$	0	–2	0
		^1E_g	$e'e'\alpha\beta,\ e''e''\alpha\beta$	0	0	2
^1E_g	^1E		$\frac{1}{\sqrt{4}}\,[(e'e'-e''e'')\alpha\beta \pm e'e''(\alpha\beta-\beta\alpha)]$	$\frac{1}{2}$	0	0
		E_1''	$\frac{1}{3}\,(^1A_g + 2\,^1E_g)$	$\frac{1}{3}$	$-\frac{2}{3}$	0
		E_2''	$\frac{1}{6}\,(^1A_g + 2\,^1E_g + 3\,^3A_g)$	$\frac{2}{3}$	$\frac{2}{3}$	0

Note: e' and e'' are the partner orbitals of 2D representations in Slater determinants, and α and β are the basis spin functions.

which apparently turns out to be the ultimate accuracy of this method as far as the system of dangling bonds in diamond is concerned. In calculations by the density functional method, the systematic error of correlation energy has the opposite sign, which is the main cause of the discrepancy between calculated $\Delta_{st} \approx 1.2$ eV shown in Fig. 2 and the value 0.25 eV found by the density functional theory [6]. Other causes are related to the parametrization and the specific features of the model; in particular, the representation of band structure disturbed by a point defect in terms of the cluster model [6] differs from that in terms of the LUC model used in this study.

Both approaches lead us to the conclusion that the total spin of the ground state is $S = 1$; however, the actual value of Δ_{st} can be estimated only from an experiment, such as that proposed below. However, the above discussion has still left unspecified the main difference between the restricted open-shell Hartree–Fock method ROHF and the calculations with violation of spin symmetry, when the many-electron state is not an eigenfunction of the squared total spin, of which the polarized-spin functional method [6] is an example.

REBONDING EFFECT

Even at a small off-center shift of a Si atom, the energy of the spin–triplet term exerts a stepwise decrease by 1.3 eV, as is shown by an arrow in Fig. 2; in contrast, no step in the energy of a singlet terms is observed. Such an "instability" of the fully symmetric atomic configuration can be attributed neither to the Jahn–Teller effect (term $^3A_{2g}$ is orbitally nondegenerate) nor to a change in the type of the open shell filling [12, 13]. The actual reason lies with the chemical-bond switching (rebonding) effect [11], which may be explained in the context of the open-shell model.

In a fully symmetric atomic configuration, it is only the a_{1g}, a_{2u}, and e_u molecular orbitals that include a contribution from Si atomic orbitals and bind the atom to the diamond. Even a small off-center shift of a Si atom $\mathbf{C}_{3v}$ allows the involvement of Si AOs (3*s*, 3*p*) in any binding MOs except a_2 but including the open-shell molecular orbital *e*. The electron-density maps indicate a considerable difference in the structure of the open-shell terms with different spin when Si is at an off-center position. Thus, the structure of $^3A_2$ comprises three dangling bonds of the carbon atoms from which Si moves away; in contrast, the structure of the ^1E term remains nearly what it is with the Si atom at the central position, i.e., contains all the six dangling carbon bonds. For the spin-triplet term, the allowance for the Si AOs contribution to open-shell MOs appears to be energetically favorable: as a result of transition from $^3A_{2g}$ to $^3A_2$, approximately 3% of density falls to the Si AOs after the self-consistency is settled. In contrast, in case of the spin-singlet term, the allowance for the involvement of Si AOs in the open-shell MOs for ^1E_g to ^1E transition yields no energy benefit and the self-consistency yields only a very small contribution from Si AOs to the open-shell MOs (about 0.01%). This implies the continuous, without rebonding, transformation of the MOs of the fully symmetric atomic configuration into the MOs of

the off-center configuration; accordingly, no step in the total energy is expected.[1]

For one of the terms of the ROHF multiplet, a formal explanation of the energy step accompanying the transition to the off-center position can be obtained from the single-open-shell approximation [13]. Since the involvement of Si orbitals is symmetry-forbidden in the open shell e_g^2 but is energetically favorable for the triplet term of the fully symmetric atomic configuration, the correlation correction from the electron configurations involving Si AOs, for example, from e_u^2, for the triplet term is larger than that for a singlet term. Thus, the rebonding effect for off-center impurities in semiconductors may be described not only empirically (as a dependence of separate contributions to the total energy on interatomic spacing) [11], but also via the configuration interaction by the MO method.[2] As a result, the ground state of the center is the spin triplet $^3A_2(\mathbf{C}_{3v})$ separated from the spin-singlet state by more than 2.5 eV, and the ESR spectrum KUL1 [8] may be related to the ground state of the $[\text{Si}–V]^0$ complex.

THE ATOMIC CONFIGURATION TUNNELING

The adiabatic behavior of all terms with Si in an off-center position suggests that their stabilization energies are negligibly small and, hence, may be characterized as tunneling barriers. Even at small shifts of Si (≈0.01 Å), total energy changes insignificantly and stabilizes when the relaxation of the carbon triad which is approached by Si moves to reduces by approximately this value and the relaxation of the other neighboring triad increases by the same value. Within the accuracy of quantum-chemical calculation, one may contend that the off-center position stabilizes the total energy of center in the triplet state. In addition, it is hardly possible to assert that such a shift does not stabilize the total energy of singlet states, as is shown in Fig. 2; however, whatever the case, the energy barriers are low and the corresponding shifts may be thermally activated and lead to the tunnel splitting of all of the terms.

Attempts to trace the adiabatic behavior of the spin-triplet state to the end failed, since, when the shift of the Si atom is larger than that shown in Fig. 2, the number of neighbors that are included in the lattice summation changes and the total energy undergoes a jump comparable to the small value of stabilization energy. However, the adiabatic character of the curve clearly suggests that stabilization is reached at a shift of ≈0.06 Å.

No energy stabilization was also revealed by the simulation of Jahn–Teller distortion $\mathbf{D}_{3d} \longrightarrow \mathbf{C}_{2h}$ of the singlet ^1E_g state. Like in the case of divacancy, this distortion is responsible for the dangling carbon bonds pairing.

Such a behavior is consistent with the model of dynamic tunneling [2] of the 1.68-eV center between two atomic configurations, when Si is shifted to either of the carbon triads.

DISCUSSION AND CONCLUSIONS

The most important result of this study is the findings that the ground spin-triplet state of the complex in the neutral charge state is subjected to the tunnel (not Jahn–Teller) splitting related to the off-center position of the Si atom on the trigonal axis of fully symmetric atomic configuration $\mathbf{D}_{3d}$. Because of this, the $[\text{Si}–V]^0$ complex may turn out to be the optical 1.68-eV center with the characteristic (≈1 meV) splitting of the zero-phonon line [2]. The calculated energy level of spin triplet $^3A_2$ is so much lower than the level of the spin singlet ^1E_g that the ground state of the center may be assumed to be of the spin-triplet type, despite the possible perturbations from the neighboring lattice defects, elastic stress, and the proximity of surface in CVD diamond films. Thus, the two tabulated spectra, the optical 1.68-eV spectrum and the ESR spectrum KUL1 [8] may originate from the same $[\text{Si}–V]^0$ complex.

It is not improbable that this conclusion is consistent with the results of the density functional calculation by Goss *et al*. [6], who reconed that it is sufficient to prove that the central position of Si is more favorable than the substitutional position at a vacancy site (Fig. 1a). Along with the conclusion that the ground state of the neutral complex is of spin-triplet type, another result that qualitatively correlates with this study is that the filled molecular orbitals localized at Si appear to be responsible for the intracenter 1.68-eV optical transition. Figure 1b shows the calculated main contribution to the lowest single-particle excitations of the neutral complex: the electron transport is shown with an arrow starting from the open shell e that transforms into e_g for the central position of Si and pointing to the other e shell that is highly localized at Si and transforms into e_u for the central position. The two shells that are resonant with the transition turn out to be band shells (labeled as "band"), since they are delocalized. With the involvement of these shells in the intracenter transition disregarded, the

[1] The different behavior of the singlet and the triplet terms can be qualitatively explained by the Hand's law. The electrons of a triplet term share the same space; hence, even a small addition of Si AOs to their molecular orbitals enables a greater variation of their total density and, therefore, decreases the correlation energy. The attainment of self-consistency results in the Si AOs transfer from the valence MOs into the open shell, which process appears as the rebonding effect. Unlike this situation, the electrons of a singlet term are spatially separated, their density in the center is lowest, and the introduction of additional Si AOs gives no gain in the correlation energy.

[2] The rebonding effect is not only a feature in the behavior of an impurity atom but also an indication for the calculation method: it signifies that the correlation corrections to different terms are unequal. Clearly, this indication is present only if both the spin and the orbital symmetry are retained throughout the calculation and is absent if the spin-polarized density functional with the self-coordination procedure common for all terms is employed [6].

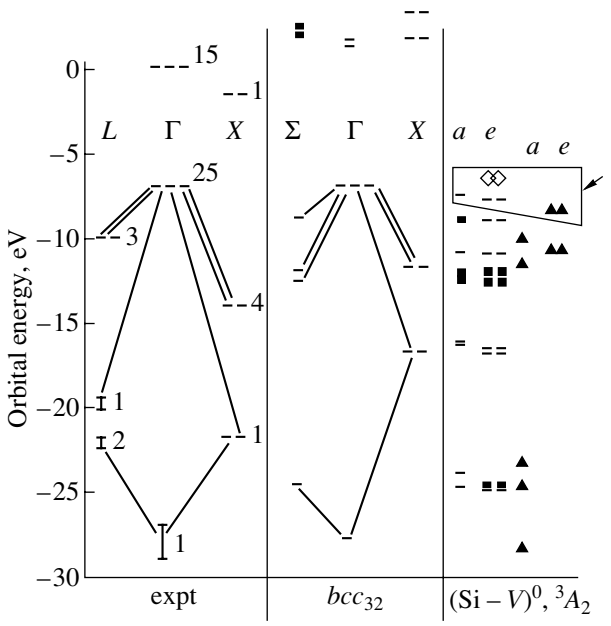


Fig. 3. One-electron spectra of the center and the defect-free crystal: expt, experimental data for diamond; bcc_{32}, calculation in terms of 32-atom LUC $\{\Gamma + 12\Sigma + 3X\}$ of diamond [12]. Lines connect energy values in the same valence band; triangles mark the Si-to-crystal bonding orbitals; and rhombuses show the open-shell orbitals. Enclosed orbitals are those corresponding to the scheme in Fig. 1b.

transition energy can be estimated from the difference between orbital energies at the arrow ends at a value of ≈1.8 eV. Due to the involvement of other states in single-electron excitation, primarily, of localized ones (triangles in Fig. 3), the energy of the transition decreases. In contrast to the case of a closed shell, the determination of the intrinsic energy of transition by the Δ*SCF* (self-consistent field) [10] method is difficult here; however, the agreement with the 1.68-eV optical spectrum is evident anyway. Optical transition with the involvement of MOs of the same origin as in this study (an open shell e_g and a closed shell e_u) has been calculated using the density functional method [6], but only for the $[Si–V]^-$ state. The same transition, ${}^2E_g \longrightarrow {}^2E_u$, was suggested to be responsible for the optical 1.68-eV spectrum (nonpolarized transition).

The scheme of the optical excitation in Fig. 1b implies that the transition is multiplet: ${}^3A_2^*$, ${}^3E^* \longrightarrow {}^3A_2$. Hence, it follows that, by measuring the polarization of 1.68-eV spectrum, one could determine its charge state. Furthermore, the lines of the center (737, 756, 767 nm, etc.) are considered as phonon replicas [4]; however, this concept fails to explain the absence of multiple overtones even for such an intensive line as 767 nm, which corresponds to the 515 cm^{-1} phonon. The presumed role of this phonon as an indicator of a Si–Si pair present in the composition of the center also lacked confirmation.

Let us finally draw some conclusions concerning the singlet terms of the neutral complex. The dipole-forbidden optical intracenter transition ${}^1E_g \longrightarrow {}^1A_{1g}$ is estimated at a value of ≈0.7 eV (Fig. 2). Optically simulated ESR or IR absorption measurements, which imply the light-induced nonequilibrium population of ${}^1E_g$ term may turn out to be a convenient means for the further investigation of the center itself. In particular, for future advances in the electron theory of defects in diamond, it is desirable to determine the value of singlet–triplet splitting in the neutral state and to reveal the possible role of a spin-singlet orbital doublet in the optical 1.68-eV transition.

ACKNOWLEDGMENTS

This study was supported by the State Scientific–Technological Program “Fullerenes and Atomic Clusters.”

REFERENCES

1. V. S. Vavilov, A. A. Gippius, and E. A. Konorova, *Electronic and Optical Properties of Diamond* (Nauka, Moscow, 1985).
2. C. D. Clark, H. Kanda, I. Kiflawi, and G. Sittas, Phys. Rev. B **51**, 16681 (1995).
3. J. Serrano, M. Cardona, and T. Ruf, Solid State Commun. **113**, 411 (2000).
4. A. A. Gorokhovsky, A. V. Turukhin, R. R. Alfano, and W. Phillips, Appl. Phys. Lett. **66** (1), 43 (1995).
5. A. V. Turukhin, C.-H. Liu, A. A. Gorokhovsky, *et al.*, Phys. Rev. B **54**, 16448 (1996).
6. J. P. Goss, B. Jones, S. J. Breuer, *et al.*, Phys. Rev. Lett. **77**, 3041 (1996).
7. K. Iakoubovskii and G. J. Adriaenssens, Phys. Rev. B **61**, 10174 (2000).
8. K. Iakoubovskii and A. Stesmans, Phys. Rev. B **66** (19), 195207 (2002).
9. S. S. Moliver, in *Book of Abstracts of the 5th Biennial International Workshop, IWFA C'2001* (FIZINTEL, St. Petersburg, 2001), p. 314.
10. S. S. Moliver, Fiz. Tverd. Tela (St. Petersburg) **41**, 404 (1999) [Phys. Solid State **41**, 362 (1999)].
11. S. T. Pantelides, W. A. Harrison, and F. Yndurain, Phys. Rev. B **34**, 6038 (1986).
12. S. S. Moliver, Fiz. Tverd. Tela (St. Petersburg) **42**, 655 (2000) [Phys. Solid State **42**, 673 (2000)].
13. S. S. Moliver, Fiz. Tverd. Tela (St. Petersburg) **42**, 1590 (2000) [Phys. Solid State **42**, 1634 (2000)].

Translated by A. Sidorova

Technical Physics, Vol. 48, No. 11, 2003, pp. 1454–1459. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 95–101.

**OPTICS,
QUANTUM ELECTRONICS**

Relation between the Spatial Growth Rate of a Traveling Wave and the Temporal Growth Rate of a Standing Wave

N. S. Bukhman and S. V. Bukhman

Samara State Architecture and Building Academy, Samara, 443001 Russia

e-mail: buhman@ssaba.smr.ru

Received October 22, 2002

Abstract—The relation between the spatial growth rate of a traveling wave and the temporal growth rate of the corresponding standing wave is examined. It is shown that, when radiation propagates in a gain medium with a sufficiently narrow gain line and a high amplification coefficient in the line center, the frequency dependences of the spatial and temporal growth rates of the field amplitude can differ significantly. In particular, at a fixed population inversion, the unbounded narrowing of the gain line, which results in an unbounded increase in the spatial growth rate and the narrowing of its frequency profile, is accompanied by neither an unbounded increase in the maximum value of the temporal growth rate nor an unbounded narrowing of the frequency profile of this growth rate. © *2003 MAIK "Nauka/Interperiodica".*

1. It is well known that the amplitude of a wave propagating in an active medium (a medium with population inversion) increases with time and/or distance. This increase is usually characterized by one of the two closely related medium parameters, namely, the spatial or temporal growth rate of a monochromatic wave. The main purpose of this study is to examine the relation between these two different parameters of the same medium. We note that this relation has a form of a trivial linear dependence only when the frequency dependence of the growth rate is fairly smooth; however, we consider the opposite case in which the growth rate is a sharp function of the frequency (a narrow gain line).

In particular, we are interested in the growth rate of the field amplitude in the cavity of a laser operating in the giant pulse mode after Q-switching. The point is that the spectral width of the laser pulse is generally believed not to significantly exceed the gain line width [1, 2]. However, it follows from the conventional theory of giant pulse generation [1, 2] (which is based on solving equations describing the growth of the field amplitude and does not take into account the finite width of the gain line) that, when the gain line width decreases without bound at a fixed population inversion or, the more so, when the lifetime of the excited state increases at a fixed pump power, which results in an increase in the equilibrium inversion, an unbounded increase in the spatial growth rate at the center of the spectral line [3] should lead to an unbounded increase in the temporal growth rate of the field in the cavity, which contradicts the above statement.

Studying the relation between the spatial and temporal growth rates allows one to examine how the above contradiction is resolved, i.e., what happens when the narrowing of the gain line (or an increase in the population inversion) leads to so large a growth rate of the field amplitude after Q-switching that the spectral width of the generated pulse should become wider than the gain line width.

In the general case, the amplitude of a plane wave that is inhomogeneous in space and time [4–6] can be represented in the form $\exp(-i\omega t + ikz)\exp(\alpha z + \beta t)$, where ω and k are the real-value frequency and wavenumber, which characterize the temporal and spatial variations in the wave phase, and α and β are the spatial and temporal growth rates, which describe the change in the wave amplitude in space and time, respectively. In a dispersive medium, the complex wavenumber $k - i\alpha$ and the complex wave frequency $\omega + i\beta$ obey a dispersion relation [4–6]

$$D(\omega + i\beta, k - i\alpha) = 0. \quad (1)$$

Relation (1) can be treated as either one complex or two real equations that reduce the number of the independent real parameters of an inhomogeneous plane wave from four to two. When the real part of the frequency ω and the real part of the wavenumber k are set independently, a plane wave satisfying the dispersion relation is inhomogeneous in both space ($\alpha(\omega, k) \neq 0$) and time ($\beta(\omega, k) \neq 0$). In this case, the wave is periodic neither in space nor in time.

By demanding (based on certain physical reasoning, e.g., assuming steady-state lasing) the wave monochromaticity (i.e., its homogeneity in time, $\beta = 0$), the number of independent parameters can be reduced to one. For instance, the wave real frequency can be regarded

as a single independent characteristic of a plane wave. Then, resolving the equation $D(\omega, k - i\alpha) = 0$ with respect to k and α, one can express the real wavenumber and the spatial growth rate through the frequency of the monochromatic wave: $k = k(\omega)$ and $\alpha = \alpha(\omega)$.

Quite similarly, when it is physically clear that the imaginary part of the wavenumber should be zero (i.e., the wave is standing[1] and periodic in space, $\alpha = 0$), one can express (by solving the equation $D(\omega + i\beta, k) = 0$) the real wavenumber and the temporal growth rate of the standing wave through its real frequency: $k = k(\omega)$ and $\beta = \beta(\omega)$.[2]

It is important that both linear gain characteristics ($\alpha(\omega)$ and $\beta(\omega)$) belong to the same medium and can be found from the same dispersion relation (1). However, they characterize the growth of the field amplitude under quite different conditions; therefore, the relation between them is generally nonlocal (with respect to frequency) and nonlinear. Namely, knowledge of the spatial growth rate $\alpha(\omega)$ and the wavenumber $k(\omega)$ at a given wave frequency ω does not allow one to find the temporal growth rate $\beta(\omega)$ for the same frequency (nonlocality with respect to frequency) and a severalfold increase in the spatial growth rate does not lead to the same increase in the temporal growth rate (nonlinearity). The point is that, strictly speaking, a standing wave that grows in time exponentially is not monochromatic.[3] Hence, its temporal growth rate depends on the properties of the medium within a certain frequency interval (nonlocality), rather than at a certain single frequency. The width of this interval increases with increasing temporal growth rate, which leads to the nonlinear dependence between the spatial growth rate of a propagating wave and the temporal growth rate of a standing wave.

Let us assume that dispersion relation (1) is resolved with respect to the complex wavenumber; i.e., it is written in the form

$$k - i\alpha = K(\omega + i\beta), \tag{2}$$

where $K(x)$ is a known function.[4]

Then, we have an explicit expression for the spatial growth rate:

$$\alpha(\omega) = -\mathrm{Im}(K(\omega)), \tag{3}$$

whereas to find the temporal growth rate, we need to solve the equation

$$\mathrm{Im}[K(\omega + i\beta)] = 0, \tag{4}$$

which determines the temporal growth rate $\beta(\omega)$ as an implicit function of the real frequency ω.

2. Let us consider Eq. (4). Assuming the parameter β to be sufficiently small, the function $K(\omega + i\beta)$ in Eq. (4) can be linearized with respect to β (see, e.g., [5]) so that Eq. (4) becomes linear with respect to this variable. The general solution to this equation has the form

$$\beta(\omega) = \alpha(\omega)v_{gr}(\omega), \quad 1/v_{gr}(\omega) \equiv \partial \mathrm{Re}K(\omega)/\partial\omega, \tag{5}$$

where $v_{gr}(\omega)$ is the conventional group velocity of a wave packet in a medium [4–6].

Sometimes, the less exact relation

$$\beta(\omega) = \alpha(\omega)v_{ph}(\omega), \quad v_{ph}(\omega) \equiv c/n(\omega) \tag{6}$$

is used instead of relation (5).

It is important that Eqs. (5) and (6) are approximate relations, which are valid only for a sufficiently low wave amplification or absorption. These relations are sometimes (see, e.g., [1, 2]) interpreted based on the identification of the wave group velocity with the energy transfer velocity. Strictly speaking, such an identification is justified only for a medium with no absorption or amplification, i.e., without both spatial and temporal growth of the wave amplitude [7].

For this reason, let us consider the relation between the spatial and temporal growth rates of a wave in an

[1] For example, when a laser operates in the giant pulse mode [1–3], the wave in the cavity after Q-switching can be considered to be standing as long as the energy gain in the cavity (due to inversion dumping) dominates over the radiative energy loss into the surrounding space. Therefore, in this case, it is the temporal growth rate of a standing wave that determines the inversion dumping time (i.e., the rise time of the giant pulse).

[2] Strictly speaking, the field in the cavity can be considered to be a superposition of plane waves with real wavenumbers only if the active medium resides in a cavity without energy losses (including the radiative loss). Thus, in the case of a one-dimensional cavity with flat mirrors (a Fabry–Perrot cavity [1, 2]), each cavity mode consists of two counterpropagating plane waves and the boundary conditions are $\cos(2kl) = 1$, where l is the cavity length. These boundary conditions can be met only when the wavenumber k is real. In laser cavities of a more sophisticated design, the boundary conditions are more complicated; however, if there is no radiative energy loss into outer space and no energy absorption in the mirrors, these conditions result in denumerable set (numbered with index n) of modes, each of which is a superposition of plane waves with a definite real wavenumber k and definite complex frequency $\omega + i\beta$. Hence, each mode (standing wave) is characterized by four parameters (ω, β, k, and n), only one of which is independent, whereas all the others are its functions. Usually, the mode index n is taken as an independent parameter and the mode index functions $\omega = \omega_n$, $\beta = \beta_n$, and $k = k_n$ ($n = 1, 2, 3, \ldots$) are considered. In this study, we will not specify the cavity geometry and will consider the field real frequency ω to be the independent parameter; hence, we will study the function $\beta(\omega)$ instead of the two functions β_n and ω_n. The advisability of such a choice stems from the universal character of the $\beta(\omega)$ function—the temporal growth rate of an arbitrary standing wave in a given medium, which is defined by Eq. (3) and does not depend on the cavity geometry. Of course, it is necessary to remember that, for any specific cavity, the field frequencies ω corresponding to the integer mode numbers n (rather than arbitrary field frequencies) are of physical meaning. The spectrum of the possible ω_n values is determined by both the active medium parameters and the cavity geometry.

[3] The wave can be quasi-monochromatic by itself (at $\beta \ll \omega$) but, at the same time, can be regarded as substantially nonmonochromatic for the given medium if $\beta \sim \Delta\Omega$, where $\Delta\Omega$ is the frequency interval over which the medium refractive index changes significantly.

[4] For example, $K(\omega) = (\omega/c)n(\omega)$, where $n(\omega)$ is the complex refractive index of the medium.

active medium when relation (6) is no longer applicable.

For the sake of definiteness, we consider the propagation of light in a medium with the refractive index $n(\omega) = n_0 + \Delta n(\omega)$ [4–6], where $\Delta n(\omega)$ is a complex additive caused by the finite width of the gain line with the central frequency ω_0 and n_0 is the background (nonresonant) refractive index, which depends slightly on the light frequency in the vicinity of ω_0. We introduce the amplitude amplification coefficient α_0 at the center of the spectral line ω_0,

$$\alpha_0 \equiv ik_0\Delta n(\omega_0), \quad k_0 \equiv \omega_0/c, \tag{7}$$

and the complex form factor of the line $g(\Omega)$, normalized to unity at the central frequency ω_0 (here, Ω is the detuning of the wave frequency from the central frequency, $\omega \equiv \omega_0 + \Omega$),

$$g(\Omega) \equiv ik\alpha_0^{-1}\Delta n(\omega_0 + \Omega). \tag{8}$$

Then, Eqs. (3) and (4) can be rewritten in the form

$$y - \gamma \mathrm{Re}[g(x + iy)] = 0 \tag{9}$$

and

$$y = \frac{\gamma \mathrm{Re}[g(x)]}{1 + \gamma \partial \mathrm{Im}[g(x)]/\partial x}, \tag{10}$$

respectively, where we have introduced the dimensionless frequency

$$x \equiv 2\Omega/\Delta\Omega_{1/2} = \Omega\tau_l, \tag{11}$$

the dimensionless temporal growth rate

$$y \equiv \beta(x)\tau_l, \tag{12}$$

and the dimensionless parameter

$$\gamma \equiv \alpha_0 c\tau_l/n_0. \tag{13}$$

In Eq. (11), $\Delta\Omega_{1/2}$ is the full width at half-maximum (FWHM) of the spectral line, τ_l is the characteristic coherence time of the spectral line ($\tau_l \equiv 2/\Delta\Omega_{1/2}$), and $g(x)$ is the complex form factor of the spectral line, whose width is normalized so that $\mathrm{Re}[g(1/2)] = 1/2$.

The parameter γ in Eq. (13) is the ratio of the temporal growth rate at the central frequency, $\alpha_0 c_l/n_0$, calculated by formula (6), to the spectral line width $\Delta\Omega_{1/2}$. Hence, this parameter reflects the degree to which a wave exponentially growing in time is nonmonochromatic with respect to the line width.

It can be seen that Eq. (10) is a consequence of Eq. (9) and an extra linearization of the function $g(x + iy)$ with respect to y in the vicinity of the point $y = 0$. Such a linearization is justified only if $y \ll 1$, or, in other words, when the temporal growth rate of a standing wave is small as compared with the spectral line width (i.e., at $\gamma \ll 1$).

For this reason, let us analyze relation (9) itself. It can be solved only numerically; hence, we first write out simpler equations for the parameters characterizing the dependence of the dimensionless temporal growth rate on the dimensionless frequency, $y(\gamma, x)$, for different values of the nonlocality parameter γ. From Eq. (9), we obtain the following equation for the dimensionless temporal growth rate $y_0 \equiv y(0)$ at the center of the spectral line:

$$y_0 - \gamma \mathrm{Re}[g(iy_0)] = 0. \tag{14}$$

For the FWHM $x_{1/2}$ of the function $y(x_{1/2}/2) = y_0/2$, we have

$$y_0/2 - \gamma \mathrm{Re}[g(x_{1/2}/2 + iy_0/2)] = 0. \tag{15}$$

3. For the Lorentz form factor of the spectral line,

$$g(x) = 1/(1 - ix). \tag{16}$$

Equations (14) and (15) can be solved analytically. In this case, instead of Eq. (9), we have

$$x^2 + (1 + y)^2 = \gamma(1 + 1/y), \tag{17}$$

and the solution to Eq. (14) is

$$y_0 = (\sqrt{1 + 4\gamma} - 1)/2, \tag{18}$$

whereas the solution to Eq. (15) is

$$x_{1/2} = (1/2)\sqrt{12\gamma + 10\sqrt{1 + 4\gamma} + 6}. \tag{19}$$

If $\gamma \ll 1$, then, instead of Eqs. (17)–(19), we obtain

$$y = \gamma/(1 + x^2), \quad y_0 = \gamma, \quad x_{1/2} = 2. \tag{20}$$

It is this (and only this) case in which the frequency dependence of the temporal growth rate $\beta(\Omega)$ of a standing wave coincides with the frequency dependence of the spatial growth rate $\alpha(\Omega)$, the maximum temporal growth rate depends linearly on the maximum spatial growth rate, and the width of the frequency dependence of the temporal growth rate coincides with the spectral line width.

In the opposite limiting case ($\gamma \gg 1$), instead of Eqs. (17)–(19), we have

$$x^2 + y^2 = \begin{cases} \gamma, & x^2 < \gamma \\ 0, & x^2 > \gamma, \end{cases} \quad y_0 = \sqrt{\gamma}, \quad x_{1/2} = \sqrt{3\gamma}. \tag{21}$$

It can be seen from these relations that, at $\gamma \gg 1$ (i.e., when either the growth rate at the center of the spectral line is high enough or the spectral line is sufficiently narrow), the frequency dependence of the temporal growth rate differs substantially from the form factor of the spectral line. In particular, instead of a Lorentz profile, we have a semicircle profile.[5] At a fixed width of the spectral line, the width of this profile increases in proportion to the square root of the spatial growth rate at the center of the spectral line. The maximum value of

[5] Remember that the first of formulas (21) describes a circle with the center in the coordinate origin and the radius $\sqrt{\gamma}$.

the temporal growth rate increases with increasing spatial growth rate in the same way.

The results obtained are shown in Fig. 1, which presents the frequency dependences of the growth rates of a standing wave for $\gamma = 0.1$, 1, and 10. The same results for the cases of a Gaussian and time-of-flight form factor of the spectral line are shown in Figs. 2 and 3, respectively. It can be seen that, in all three cases, the situation is qualitatively the same.

Thus, we can conclude that the temporal growth of the field amplitude in an active medium is governed by the parameter $\gamma \equiv \alpha_0 c\tau_l/n_0$, which is small for the low amplification coefficient and broad spectral line and large for the high amplification coefficient and narrow spectral line.

When the parameter γ is small, the frequency dependence of the temporal growth rate coincides with the line profile, the width of the frequency dependence of the temporal growth rate is equal to the spectral line width, and the maximum value of the temporal growth rate of a standing wave increases linearly with increasing the spatial growth rate at the center of the spectral line and is independent of its width.

When the parameter γ is large, the frequency dependence of the temporal growth rate of a standing wave has a characteristic semicircle shape; namely, it slightly varies near the center of the spectral line and sharply increases near certain boundary frequencies. In this case, the width of the frequency dependence of the temporal growth rate, as well as its maximum value, is proportional to the square root of the product of the spectral line width by the spatial growth rate at the line center.

4. The physical meaning of the results obtained is easy to understand taking into account that the rapid growth of the amplitude of a monochromatic wave breaks its monochromaticity. Strictly speaking, only a wave with a constant (in time) amplitude can be considered monochromatic. Hence, a wave that behaves in time as $\sim\exp[-i(\omega + i\beta)t]$ can be regarded as a wave with the carrier frequency ω and a nonzero spectral width on the order of β. This circumstance does not affect the growth of the field amplitude if the temporal growth rate of the wave is small as compared to the spectral line width ($\beta \ll \Delta\Omega_{1/2}$).[6] Otherwise ($\beta \gg \Delta\Omega_{1/2}$), if the wave grew with a growth rate exceeding the gain line width, only a small fraction of the wave spectrum would fall within the gain bandwidth; thus, the wave should grow at a much lower growth rate than that corresponding to the spatial growth rate at the center of the line. The self-consistent temporal growth rate near the line center can be estimated as $\beta = \alpha_0(c/n_0)(\Delta\Omega_{1/2}/\beta)$, where it is taken into account that, for β larger than the spectral line width $\Delta\Omega_{1/2}$, only a fraction (proportional to $\Delta\Omega_{1/2}/\beta$) of the wave spectrum falls within the spectral line. It is this fraction of the spectrum that determines the field growth rate. Thus, the temporal growth rate near the line center can be estimated as $\beta_{\max} \sim \sqrt{\alpha_0(c/n_0)\Delta\Omega_{1/2}}$, which agrees with formula (21). To estimate the width of the frequency dependence of the temporal growth rate, it is sufficient to take into account that, when $\Omega \le \beta$, the detuning of the wave frequency from the central frequency is of minor importance because the width of the wave spectrum exceeds this detuning. Consequently, the width of the frequency dependence of the growth rate $\Delta\Omega_{1/2}^{(t)} \equiv x_{1/2}/\tau_l$ is approximately equal to the maximum value of

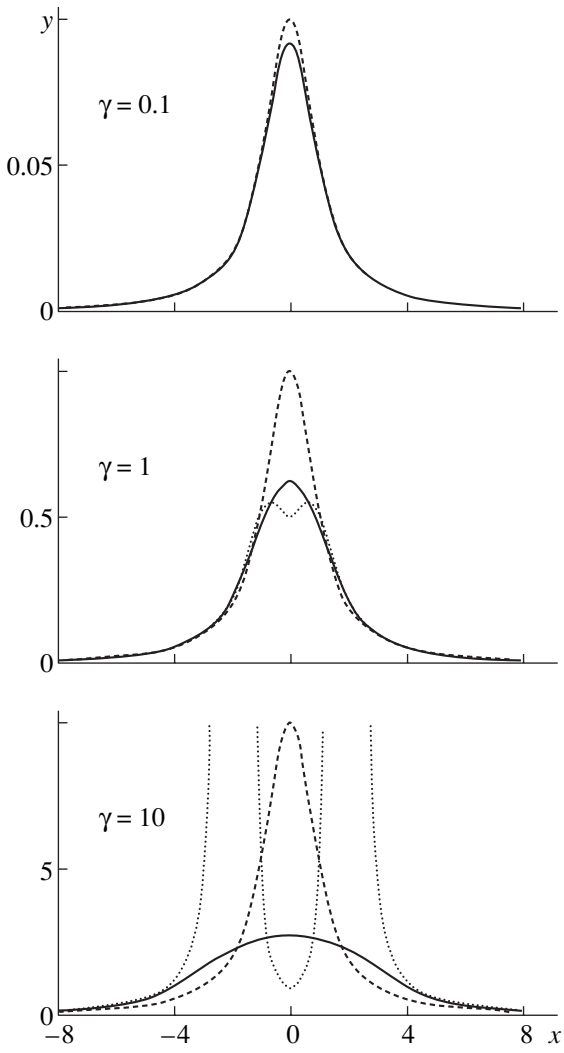


Fig. 1. Frequency dependence of the temporal growth rate for $\gamma = 0.1$, 1, and 10 and a Lorentz profile of the spectral line. The solid curves show the results calculated using Eq. (4) without any additional approximations; the dashed curves show the results calculated using the "phase-velocity approximation" (by formula (6)); and the dotted curves show the results calculated using the "group-velocity approximation" (by formula (5)).

[6] We note that, in this case, $\gamma \ll 1$.

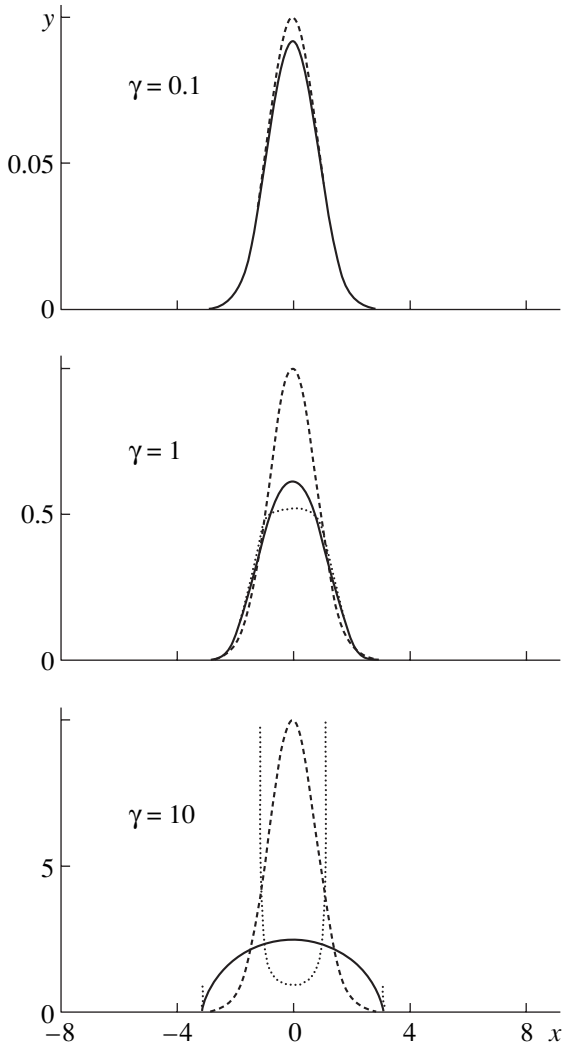


Fig. 2. Same as in Fig. 1 but for a Doppler profile of the spectral line.

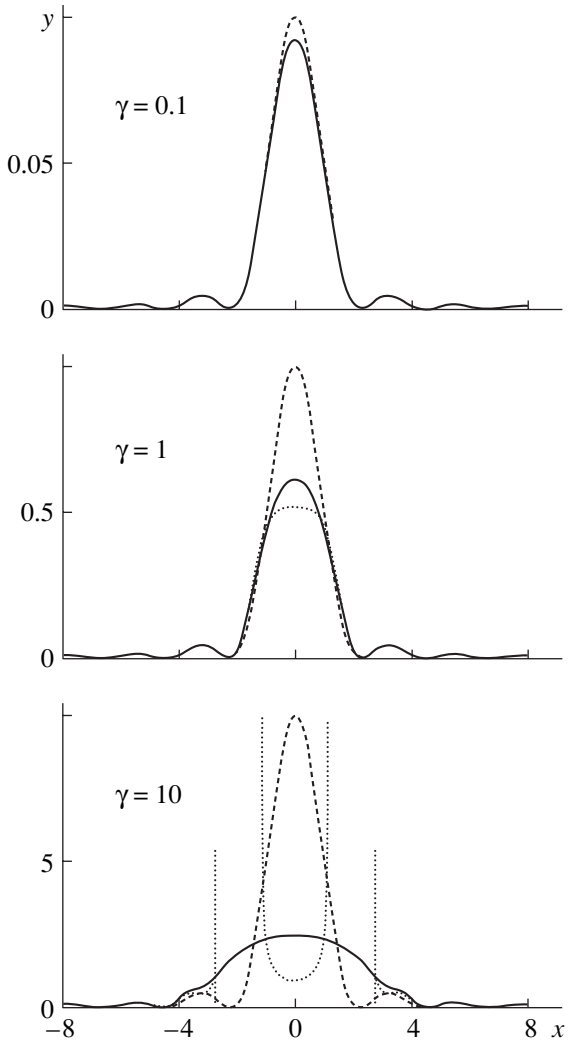


Fig. 3. Same as in Fig. 1 but for a time-of-flight profile of the spectral line.

the temporal growth rate;[7] i.e., $\Delta\Omega_{1/2}^{(t)} \sim \beta_{max}$, in accordance with formula (21).

The sharp increase in the temporal growth rate at the boundaries of the amplification band for $\gamma \gg 1$ is related to a trigger nature of the broadening of the spectrum of a monochromatic wave when its amplitude exponentially grows in time. If, for a certain "seed" growth rate of a wave, the broadening of the wave spectrum leads to an increase in the growth rate, then it will cause the further broadening of the spectrum and the further increase in the growth rate[8] and so on, until the wave spectrum covers the central region of the spectral line. After that, the broadening of the wave spectrum leads to a decrease (rather than an increase) in the growth rate and, thus, comes to an end.

It should be noted that, in the case of an unbounded narrowing of the spectral line ($\Delta\Omega_{1/2} \longrightarrow 0$) and the corresponding increase [1–3] in the amplification coefficient at the center of the spectral line ($\alpha_0 \longrightarrow \infty$) (at a fixed inversion, we have $\alpha_0\Delta\Omega_{1/2}$ = const), the frequency dependence $\beta(\Omega)$ of the temporal growth rate of a standing wave significantly differs from the frequency dependence $\alpha(\Omega)$ of the spatial growth rate of a traveling wave. In this case, the temporal growth rate at the line center, β_{max}, and the width of the frequency dependence of this growth rate, $\Delta\Omega_{1/2}^{(t)}$, tend to certain finite values that depend on the inversion level attained and the transition parameters and do not depend on the

[7] Actually, there are two different frequency dependences: the frequency dependence of the spatial growth rate $\alpha(\Omega)$ and the frequency dependence of the temporal growth rate $\beta(\Omega)$. Therefore, there are two different spectral line widths, $\Delta\Omega_{1/2}$ and $\Delta\Omega_{1/2}^{(t)}$.

[8] Of course, this is only a hypothetical iterative procedure rather than a real increase in time.

spectral line width.[9] For example, for a Lorentz form factor and $\gamma \gg 1$, we find, according to the formulas for the amplification coefficient from [3], that

$$\beta_{\max} = \sqrt{(2\pi\omega_0 |d_0^{(e)}|^2 \Delta N)/(n_0 \hbar)},$$
$$\Delta\Omega_{1/2}^{(t)} = \sqrt{(6\pi\omega_0 |d_0^{(e)}|^2 \Delta N)/(n_0 \hbar)}, \quad (22)$$

where $d_0^{(e)}$ is the transition matrix element corresponding to the spectral line under study and ΔN is the inversion.

Note that, when the spectral line is sufficiently narrow and intense ($\gamma \gg 1$), neither the maximum value of the temporal growth rate nor the width of the frequency dependence of this growth rate depend on the spectral line width. The line width affects only the parameter γ, which determines the applicability range of relations (22). However, there is a certain restriction—the inversion ΔN cannot exceed the atomic density in the medium; therefore, there is a certain limiting growth rate. Thus, an unbounded narrowing of the spectral line at a fixed inversion leads to neither an unbounded increase in the maximum growth rate $\beta_{\max}$ nor an unbounded decrease in the width of the frequency dependence $\Delta\Omega_{1/2}^{(t)}$ of this growth rate. The narrowing of the spectral line at a constant pump power leads to an increase in the inversion, which results in an increase in the temporal growth rate and the spectral width of the generated pulse.

The result obtained can be interpreted as a specific version of the time-of-flight mechanism for the spectral line broadening [3]. Here, the limited time of the interaction of radiation with matter (on the order of $1/\beta$) stems from the rapid growth of the field amplitude with the rate β. In turn, the growth rate β depends on this interaction time, which results in the square-root dependences of the amplification bandwidth and the maximum value of the temporal growth rate on the inversion.

In some widely used laser systems [1–3], the parameter γ is on the order of unity[10] and, hence, the above mechanism for the spectral line broadening should be taken into consideration when analyzing the time during which a steady-state lasing is established, as well as when examining nonsteady lasing.

The results obtained can be summarized as follows:

(i) Generation of giant laser pulses is feasible for a single-mode operating mode because the nonsteady nature of the laser output signal in the regime of giant pulse generation is not related to the interference of several quasi-monochromatic modes (as is the case of mode-locking regime) but stems from the inherent unsteadiness of an individual standing wave in a gain medium.

(ii) The spectral width (and, consequently, the inverse rise time) of a giant pulse can significantly exceed not only the inverse lifetime of photons in the cavity (which fact was noted in [1, 2]) but also the width of the gain line. When the inversion is sufficiently high, the spectral width of the giant pulse does not depend on the gain line width and is proportional to the square root of the inversion reached up to the instant of Q-switching (or, in other words, the spectral width of the giant pulse is proportional to the square root of the product of the spatial growth rate by the spectral line width).

(iii) At a constant pump power, the increase in the lifetime of the upper laser level leads to an increase in the inversion. Hence, under these conditions, the narrowing of the gain line width will finally lead to an increase (rather than a decrease) in the spectral width of a giant pulse: this width turns out to be inversely proportional to the gain line width. Certainly, to obtain a giant pulse whose rise time is small compared with the inverse spectral line width, one needs to enable a sufficiently high cavity Q factor not only within the narrow spectral width of the gain line but within a wider spectral width of the giant pulse.

[9] We emphasize that the spectral line width itself (defined as a width of the $\alpha(\omega)$ function) does not vanish at all, and the same medium continues to be "a narrowband amplifier" of a small signal. Nevertheless, the spectral width $\Delta\Omega_{1/2}^{(t)}$ of the pulse generated by this medium turns out to be much larger than the spectral width of the gain line $\Delta\Omega_{1/2}$. This situation is quite natural because these two widths are associated with very different processes that can occur in the same medium, depending on the external conditions. The spectral width $\Delta\Omega_{1/2}$ is related to the amplification of a traveling wave that grows in space and, thus, can be steady in time (i.e., monochromatic). The spectral width $\Delta\Omega_{1/2}^{(t)}$ is related to a standing wave that retains its spatial distribution during amplification and, thus, varies in time; i.e., it is nonsteady (nonmonochromatic).

[10] For instance, in a low-pressure CO_2 laser with $\lambda = 10.6$ μm, $2a_0 = 4$ dB/m, and $\Delta\nu_D = 50$ MHz, we have $\gamma = 0.88$; in a He–Ne laser with $\lambda = 3.39$ μm, $2a_0 = 20$ dB/m, and $\Delta\nu_D = 280$ MHz, we have $\gamma = 0.79$; and in a YAG-laser with $\lambda = 1.06$ μm, $\alpha_0 = 20$ cm^{-1}, $\Delta\nu = 6$ cm^{-1}, and $n_0 = 1.82$, we have $\gamma = 0.58$.

REFERENCES

1. N. V. Karlov, *Lectures on Quantum Electronics* (Nauka, Moscow, 1983; CRC, Boca Raton, 1993).
2. A. Yariv, *Introduction to Optical Electronics* (Rinehart and Winston, New York, 1971; Vysshaya Shkola, Moscow, 1983).
3. D. N. Klyshko, *Principles of Quantum Electronics* (Nauka, Moscow, 1986).
4. M. B. Vinogradova, O. V. Rudenko, and A. P. Sukhorukov, *Theory of Waves* (Nauka, Moscow, 1979).
5. A. F. Aleksandrov, L. S. Bogdankevich, and A. A. Rukhadze, *Principles of Plasma Electrodynamics* (Vysshaya Shkola, Moscow, 1988; Springer-Verlag, Berlin, 1984).
6. A. I. Akhiezer and I. A. Akhiezer, *Electromagnetism and Electromagnetic Waves* (Vysshaya Shkola, Moscow, 1985).
7. M. I. Rabinovich and D. I. Trubetskov, *Oscillations and Waves in Linear and Nonlinear Systems* (Nauka, Moscow, 1984; Kluwer, Dordrecht, 1989).

Translated by N. Ustinovskiĭ

Technical Physics, Vol. 48, No. 11, 2003, pp. 1460–1463. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 102–105.

OPTICS, QUANTUM ELECTRONICS

Some Issues Concerning the Development of a Speckle Velocimeter

A. Aliverdiev[1, 2], M. Caponero[3], and C. Moriconi[1]

[1]*Robotics and Informatics Division, ENEA Casaccia, via Anguillarese 301, 00060 Roma, Italy*

[2]*Institute of Physics, Daghestan Research Center, Russian Academy of Sciences, ul. M. Yaraskogo 94, Makhachkala, 376003 Russia*

[3]*Applied Physics Division, ENEA Frascati, via Enrico Fermi 455, 00044 Frascati, Italy*

e-mail: aliverdi@mai.ru

Received February 21, 2003; in final form, May 5, 2003

Abstract—The application of integrated statistics in the space–time plane to laser speckle velocimetry is considered. A new approach to determining the contrast function is proposed. This approach makes it possible to considerably improve the stability of the solution to the problem under nonideal conditions. The results of experimental investigations are presented. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The objective of the present study was to develop a precision velocimeter for an independent robotized vehicle ("snow cat") under the conditions of the Antarctic region. Interest in this problem was generated by difficulties encountered in attempts at precisely measuring velocities by standard methods under the conditions of a probable skid and the absence of unambiguous reference points over extensive snow or ice fields in the Antarctic region. The choice of speckle velocimetry was motivated by its high precision and by the opportunity of operation under these conditions.

By the present time, we have performed a series of theoretical and experimental investigations, the results of these being quite good and promising.

HISTORY OF THE USE OF THE SPECKLE EFFECT IN VELOCIMETRY

The use of the speckle effect in velocimetry has been repeatedly discussed in the literature. Various approaches were proposed that were based on invoking either time or spatial statistics for both the time-differentiated and the time-integrated intensity function.

In 1980, Fercher [1] proposed employing the time statistics of a time-differentiated function. This made it possible to obtain quite accurate experimental results in measuring the velocity of a Plexiglas sample, the measured velocities ranging between 1 μm/s and 1 mm/s. The formulas given in Fercher's article that relate the sought velocity to statistical quantities determined experimentally are quite simple and clear.

A different approach that employed the spatial statistics of the time-integrated intensity function was described in [2], where the root-mean-square spatial deviation $\sigma_s(T)$ of the speckle intensity function over the averaging (exposure) time T was related to the averaging time and the characteristic "correlation time" by the simple equation

$$\sigma_s/\langle I\rangle = \sqrt{(\tau_c/2T)(1-e^{-2T/\tau_c})}, \qquad (1)$$

where I is the intensity of the speckle picture.

Here, the correlation time was a quantity that directly determined the velocity of motion of scatterers that generated speckles. In more recent studies, the method was refined both theoretically and experimentally [3–6].

The simplicity and clarity of information derived by this method is its basic advantage. With the aid of the method in question, one can straightforwardly estimate the spatial distribution of the velocities of scatterers. Moreover, the mathematical formalism underlying the method made it possible to achieve a rather high precision. Yet another feature inherent in this method was that it measures velocity irrespective of direction—more specifically, a calculation of the velocity on this basis provides no way either to determine its direction or, in general, to distinguish between a translational and a vibrational motion. Thus, the procedure described above found some successful applications—for example, in medical diagnostics—but it is hardly optimal in our problem.

In [7, 8], we therefore proposed somewhat modifying an analysis of the integrated statistics of the spatial and time dependence of the recorded speckle-picture intensity $I(x, t)$, namely, we proposed to perform, instead of set of averagings for various periods T, the averaging for a set of angles in the space–time plane at a single value of T. Previously, a similar approach was successfully used to interpret the experimental spatial

1063-7842/03/4811-1460$24.00

and time dependence of the intensity of spontaneous radiation associated with an electric breakdown [9, 10].

In this way, we obtained a transformation that is similar to the Radon transformation [9], but which involves fixed limits in a quasiuniform space–time plane; that is,

$$g(s, \phi) = \int_{-Tv_0}^{Tv_0} I\left(s\cos(\phi) - p\sin(\phi), \frac{s}{v_0}\sin(\phi) + \frac{p}{v_0}\cos(\phi)\right)dp, \quad (2)$$

where (s, ϕ) are the normal coordinates of the Radon transformation, p is a variable with respect to which integration is performed, and v_0 is the ratio of the time resolution of the measuring system used to its spatial resolution.

If time is measured in the number of frames and if x is measured in the number of pixels (as is shown in Fig. 1), then $v_0 = 1$; nevertheless, we must take this coefficient into account in eventual calculations.

In the case being considered, the parameter s appears as an analog of a spatial parameter that can be used in evaluating the spatial root-mean-square deviation of the speckle-intensity function; that is,

$$\sigma_s^2(\phi) = \langle (g(s, \phi) - \langle g(s, \phi)\rangle_s)^2\rangle_s. \quad (3)$$

The maximum value of the spatial root-mean-square deviation of the speckle-intensity function must be observed at the angle that corresponds to the sought velocity of the object being considered:

$$v = v_0\tan\left(\phi((\sigma_s)_{\max}) - \frac{\pi}{2}\right). \quad (4)$$

In [7, 8], we reported the results of our investigations where actual motion was simulated by finite fixed steps that were recorded on individual frames. Under these idealized conditions, the method yielded quite promising results.

FORMULATION OF THE PROBLEM

Developing the proposed approach, we addressed the problem of determining not only the absolute values of velocities but also the direction of motion—in other words, the vector $v = (v_x, v_y)$.

The creation of an automated working facility that is able to measure the two-dimensional velocity of a moving object was yet another important problem.

An implementation of steady and precision velocimetry under nonideal conditions was the final objective of the present study.

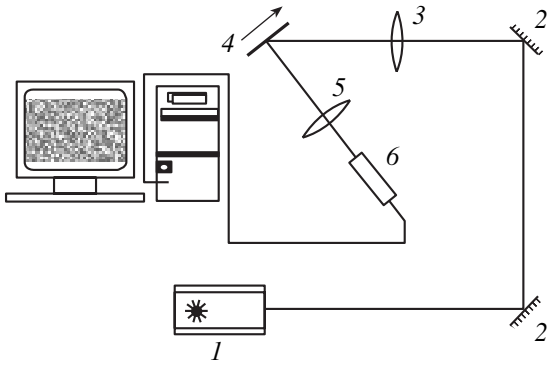


Fig. 1. Layout of the experimental facility.

DESCRIPTION OF THE EXPERIMENTAL FACILITY

The layout of the experimental facility that we used to perform our tests is depicted in Fig. 1. A beam from a Siemens LGK 7621 (Q 4001-K7 654) helium–neon laser (*1*) was directed through a system of mirrors (*2*) and lenses (*3*) to a moving object (*4*), which was taken in the form of a white-paper screen on which there was a weakly contrasted picture imitating an actual surface. The translation motion of the screen was ensured by electric motors.

After that, light scattered by the screen was recorded through a system of lenses (*5*) by a JVC TK-S350EG camera (*6*) connected through an IMAQ PCI-1408 analog-to-digital converter to an IBM-compatible computer. The visualization, saving, and primary automation of the experiment were performed by an original software package in C++, the IMAQ LabVIEW libraries (National Instruments) being used in this package.

BASIC RESULTS AND THEIR DISCUSSION

Since, in the presence of an orthogonal velocity component, the traces of speckles quickly leave the region under study, we propose performing integration in (2) over the orthogonal direction. In this way, we find for the x and y components of the velocity that

$$g_x(s, \phi) = \int_{-Tv_0}^{Tv_0}\int_{-Tv_0}^{Tv_0} I_x\left(s\cos(\phi) - p\sin(\phi), \frac{s}{v_0}\sin(\phi) + \frac{p}{v_0}\cos(\phi)\right)dpdy, \quad (5)$$

$$g_y(s, \phi) = \int_{-Tv_0}^{Tv_0}\int_{-Tv_0}^{Tv_0} I_y\left(s\cos(\phi) - p\sin(\phi), \frac{s}{v_0}\sin(\phi) + \frac{p}{v_0}\cos(\phi)\right)dpdx, \quad (6)$$

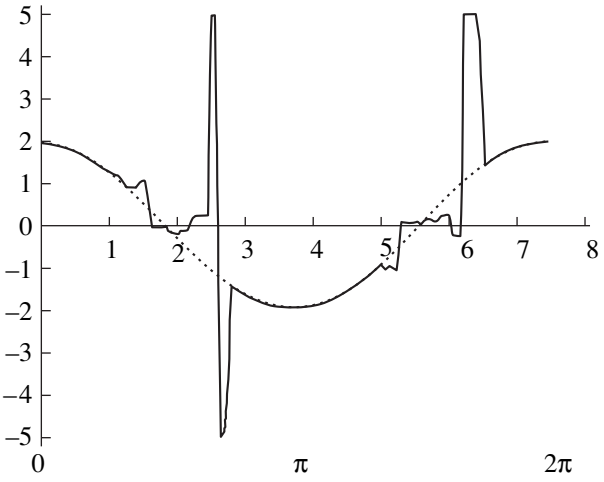


Fig. 2. Velocity as a function of the angle α of chamber rotation (that is, the angle between the *x* axis and the direction along which the velocity is measured): (solid curve) velocity (*V*) calculated on the basis of the procedure described in the main body of the text and (dotted curve) actual velocity values (*V*).

where $I_x(y, t)$ and $I_y(x, t)$ are the coordinate- and time-dependent intensities of the recorded speckle picture for, respectively, the *x* and the *y* component of the velocity.

A further treatment of the signal was performed by using Eqs. (3) and (4) for each direction.

Even the first experiments revealed, however, that, if use is made of the procedure proposed previously, a precision determination of the velocity is possible only in directions close to the true direction of the motion.

In this connection, we have examined velocity as a function of the angle α of chamber rotation—that is, the angle between the *x* axis and the direction along which velocities are measured. For each angle α, we thereby obtained a local system of coordinates (*x*', *y*') that was rotated through the angle α about the system of coordinates (*x*, *y*) and for which we calculated the velocity in the *x*' direction.

A typical result is given in Fig. 2. It can be seen that a precise calculation of velocities (with a relative error less than 2%) was indeed accomplished only within rather narrow segments close the true direction of the motion.

In order to analyze this effect, we address Fig. 3, which displays the maximum of σ_s^2 as a function of ϕ (vertical axis) and the angle α of rotation (horizontal axis). It can be seen that, although the line associated with the sought velocity can be traced rather well, it does not yield an absolute maximum within segments where there are significant errors.

It follows that, in principle, a precision 2*D* velocimetry on the basis of recorded experimental data is possible, but this would call for evaluating velocities for a wide set of angles; in turn, this would require overly great computational powers for real-time applications in practice.

In this connection, we addressed the question of revealing the factors responsible for the emergence of maxima not associated with the velocity of the motion. An analysis of the dependence $\sigma_s^2(\phi, \alpha)$ showed that these maxima are due to low-frequency components of the spectrum of the function *g*. In view of this, we proposed determining the sought value of ϕ not on the

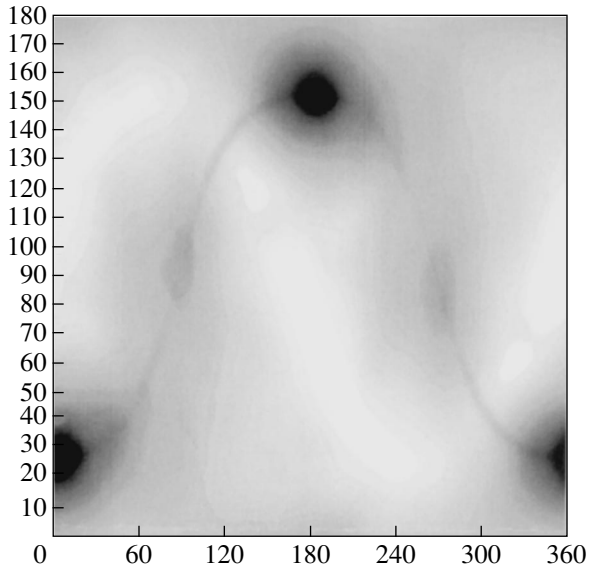


Fig. 3. Maximum of σ_s^2 as a function of ϕ (vertical axis) and the angle α of rotation (horizontal axis).

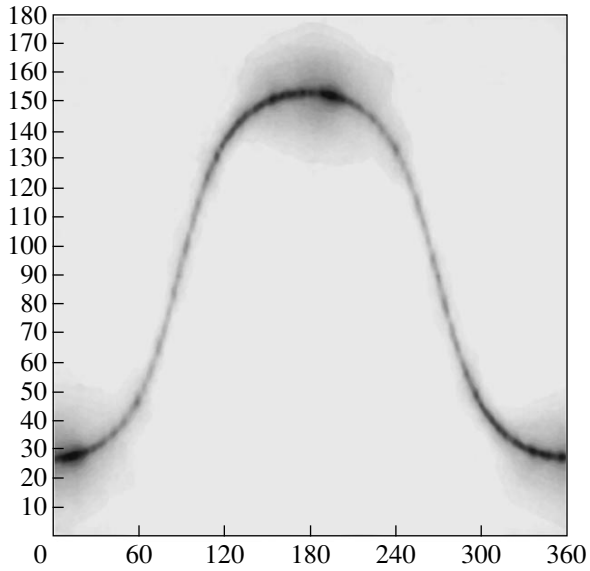


Fig. 4. Maximum of $\sigma_s'^2$ as a function of ϕ (vertical axis) and the angle α of rotation (horizontal axis).

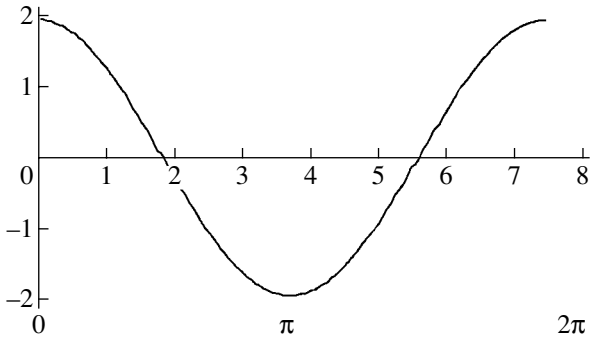


Fig. 5. Velocity calculated on the basis of the modified algorithm as a function of the angle α of chamber rotation (that is, the angle between the *x* axis and the direction along which the velocity is measured).

basis of the root-mean-square deviation in (3) but on the basis of the function calculated by the formula

$$\sigma_s'^2(\phi) = \langle (g(s + \Delta s, \phi) - g(s, \phi))^2 \rangle_s. \qquad (7)$$

The function $\sigma_s'^2(\phi)$ greatly depends on the parameter Δs. For $\Delta s \longrightarrow 0$, it also tends to zero. At a nonzero value of Δs, however, it enables us to cut off low frequencies.

For the experiment described above, the results obtained by calculating the maximum of $\sigma_s'^2$ as a function of ϕ (vertical axis) and the angle α of rotation (horizontal axis), as well as the velocity as a function of the angle α, by using this modified algorithm are illustrated in Figs. 4 and 5.

It can be seen from Fig. 5 that, for any angle, a determination of the velocity to a high precision ($\varepsilon < 2\%$) is fitted to a sine curve. In order to find the vector $\mathbf{v} = (v_x, v_y)$, we can restrict ourselves to two calculations of the velocity in orthogonal directions within any angular segment; but this was impossible without going beyond the previous algorithm.

It should be emphasized that the results presented in Figs. 2–5 are typical and were reproduced in all experiments where we changed the measured velocity and the surface of the screen.

CONCLUSIONS

In contrast to algorithms applied previously, the proposed algorithm for determining velocities from experimental data makes it possible to improve significantly the accuracy of measurements under nonideal conditions and to reduce the required computational powers considerably, the latter being of importance for real-time applications of the procedure. Thus, our present results are of great methodological value and are of importance for applications.

A detailed mathematical validation of formula (7) and the construction of a mobile velocimeter are presently under way. The relevant results will be published later.

ACKNOWLEDGMENTS

This work was supported by ENEA, the Italian Program of Researches in Antarctica (project no. 11-15), and the Ministry for Higher Education of the Russian Federation (project no. PD 02-1.2-47).

REFERENCES

1. A. F. Fercher, Opt. Commun. **33**, 129 (1980).
2. A. F. Fercher and J. D. Briers, Opt. Commun. **37**, 326 (1981).
3. A. F. Fercher, M. Peukert, and E. Roth, Opt. Eng. **25**, 731 (1986).
4. J. D. Briers and S. Webster, Opt. Commun. **116**, 36 (1955).
5. G. Richards and J. D. Briers, Proc. SPIE **2981**, 160 (1997).
6. J. D. Briers, G. Richards, and X. W. He, J. Biomed. Opt. **4**, 164 (1999).
7. M. Caponero, C. Moriconi, and A. Aliverdiev, in *Technical Digest of the 4th Italian–Russian Laser Symposium, St. Petersburg, 2001*, p. 122.
8. A. Aliverdiev, M. Caponero, and C. Moriconi, Zh. Tekh. Fiz. **72** (8), 116 (2002) [Tech. Phys. **47**, 1044 (2002)].
9. A. A. Aliverdiev, in *Proceedings of the 25th International Conference on Phenomena in Ionized Gases, Nagoya, Japan, 2001*, Vol. 4, pp. 59–60.
10. A. A. Aliverdiev, Zh. Tekh. Fiz. **67**, 132 (1997) [Tech. Phys. **42**, 1102 (1997)].

Translated by A. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1464–1467. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 106–110.

OPTICS, QUANTUM ELECTRONICS

Holographic Interferometry for Measuring Residual Stresses by Using Probing Holes

I. N. Odintsev, V. P. Shchepinov, and A. Yu. Shchikanov

Moscow State Engineering Physics Institute, Kashirskoe sh. 31, Moscow, 115049 Russia

e-mail: postgrad@postgrad.mephi.ru

Received March 25, 2003

Abstract—An original method of determining residual stresses by using probing holes and measuring the difference in the holographic interference fringe orders for two sets of pairs of points taken on the principal strain axes is suggested. The optical scheme of the interferometer is based on the use of reflection holograms. The principal residual strains are found by solving an overdetermined set of linear equations. The effect of rigid displacement on the fringe pattern is taken into account. The method is experimentally verified by measuring elastic stresses in uniaxially and biaxially strained specimens. © *2003 MAIK "Nauka/Interperiodica".*

The method of probing holes finds extensive application today for measuring residual stresses in structure elements [1]. Small holes are made in an area of interest on the surface to record a strain response of the material. Subsequent analysis consists in solving the inverse problem in mechanics of rigid body, i.e., in recovering initial residual stresses (which are zero on the free surface) from strains measured.

The local strain parameters near a probing hole are usually measured with an array of small-base resistance strain gauges. However, input information thus obtained is often limited and excludes comprehensive interpretation in terms of residual stresses. One cannot invoke statistical methods in this case and is unable to estimate the adequacy of a given mechanical model to the real strain pattern.

It seems reasonable therefore to determine residual stresses by field methods that measure strains or displacements. Such methods provide a virtually unlimited body of information. An example of field methods is the application of optically sensitive coatings [2]. However, this method, like the strain gauging method, requires special surface conditioning (application of the coating). A promising alternative in this respect is the contactless high-sensitivity method of holographic interferometry, for which surface conditioning is unnecessary.

In early works, the method of holographic interferometry was used mainly to measure displacements that are normal to the surface [3–5]. In such a configuration, information on the isotropic (spherical) stress tensor is lost, which is undesirable in most cases. Later, it was suggested that residual stresses be determined from the tangential components of displacements of points lying immediately on the hole circumference [6]. Ideally, the absolute displacements here are maximal; hence, so is the sensitivity of the method. Actually, however, the circumference experiences the highest destructive forces during drilling, which may distort the interference pattern up to its complete breakdown.

In this work, we consider a method where the difference in the holographic interference fringe orders for two sets of pairs of points taken on the principal strain axes some distance away from the hole is used as primary information. Residual stresses in this case are found by solving an overdetermined set of linear equations. Experimental data for elastic stresses in beams and thick-walled tubes support the efficiency of our method.

Holographic interferograms visualizing the 3D deformation of the surface around a probing hole were recorded according to the procedure described in [6]. It should be recalled that double-exposed holograms are recorded with an optical scheme using intersecting beams and collimated illumination normal to the surface. A photoplate is placed in a special device that makes it possible to remove the photoplate from the interferometer scheme after the first exposure and return to its position (after a probing hole has been made) with a high precision.

At the first stage of reconstruction of double-exposed holograms, the vector of observation is aligned with the normal to the object's surface (the axis x_3 of the Cartesian laboratory coordinate system $Ox_1x_2x_3$). In this case, interference fringes are the loci of points of equal displacements W from the surface. If the hole diameter is small, the local stressed state may be assumed to be uniform with a fairly high accuracy; then, the field W has two axes of symmetry, which coincide with the axes of principal residual strains (the axes x_1 and x_2 of the coordinate system). Once the principal axes have been visually set, two pairs of interferograms are recorded with vectors of observation oriented independently in the planes Ox_1x_3 and Ox_2x_3. Points in the half-space of observation are usually specified in the spherical coor-

1063-7842/03/4811-1464$24.00

dinate system (r, ψ, φ), where r is polar radius, ψ is latitude, and φ is longitude. With $r \longrightarrow \infty$, the observation of the probing hole is collimated. On the reconstructed holographic image, the center of this system coincides with the center of the hole on the surface. These pairs of interferograms correspond to the direction angles of the vector of observation $(0, \psi_1)$, (π, ψ_2) and $(\pi/2, \psi_3)$, $(3\pi/2, \psi_4)$. The optical scheme of illumination and observation of points *1* and *2* on the surface in the plane Ox_1x_3 is shown in Fig. 1.

After a probing hole has been made, each point on the surface of a test object is displaced along the principal axes x_1 and x_2 by $D(x_1) = [U(x_1), 0, W(x_1)]$ and $D(x_2) = [0, V(x_2), W(x_2)]$, respectively. Let a pair of points with the coordinates $(x_{1i}, 0, 0)$ and $(x_{1j}, 0, 0)$ that lie on the principal axis x_1 be observed at an angle ψ_k ($i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, K$). From the basic relationship of holographic interferometry [7], we obtain for either of the points (Fig. 1) two equations

$$U(x_{1i})\sin\psi_k + W(x_{1i})(1 + \cos\psi_k) = \lambda N(x_{1i}), \quad (1)$$

$$U(x_{1j})\sin\psi_k + W(x_{1j})(1 + \cos\psi_k) = \lambda N(x_{1j}), \quad (2)$$

where λ is the laser radiation wavelength and $N(x_{1i})$ and $N(x_{1j})$ are the absolute orders of fringes at the points with the coordinates $(x_{1j}, 0, 0)$ and $(x_{1j}, 0, 0)$, respectively.

Subtracting Eq. (2) from Eq. (1) yields

$$\begin{aligned}&(U(x_{1i}) - U(x_{1j}))\sin\psi_k \\ &+ (W(x_{1i}) - W(x_{1j}))(1 + \cos\psi_k) = \lambda(N(x_{1i}) - N(x_{1j})).\end{aligned} \quad (3)$$

Equation (3) can be recast as

$$(\Delta U_{1ij})\sin\psi_k + (\Delta W_{1ij})(1 + \cos\psi_k) = \lambda\Delta N(x_{1ij}), \quad (4)$$

where (ΔU_{1ij}) and (ΔW_{1ij}) are the differences between the components of the displacements U and W and $\Delta N(x_{1ij})$ is the difference in the absolute orders of fringes at the points with the coordinates $(x_{1i}, 0, 0)$ and $(x_{1j}, 0, 0)$.

An equation for points with the coordinates $(0, x_{2m}, 0)$ and $(0, x_{2n}, 0)$ that lie on the principal axis x_2 can be obtained in a similar way:

$$(\Delta U_{2mn})\sin\psi_1 + (\Delta W_{2mn})(1 + \cos\psi_1) = \lambda\Delta N(x_{2mn}), \quad (5)$$

where $m = 1, 2, \ldots, M$ and $n = 1, 2, \ldots, N$.

Generally, the functions $U(x_1)$, $W(x_1)$ and $V(x_2)$, $W(x_2)$ are the sums of the strain-induced local displacements $u(x_1)$, $w(x_1)$ and $v(x_2)$, $w(x_2)$ and generalized rigid displacements U^0, V^0, and W^0 relative to the recording medium:

$$\begin{aligned}U(x_1) &= u(x_1) + U^0; \quad W(x_1) = w(x_1) + W^0; \\ V(x_2) &= v(x_2) + V^0; \quad W(x_2) = w(x_2) + W^0.\end{aligned} \quad (6)$$

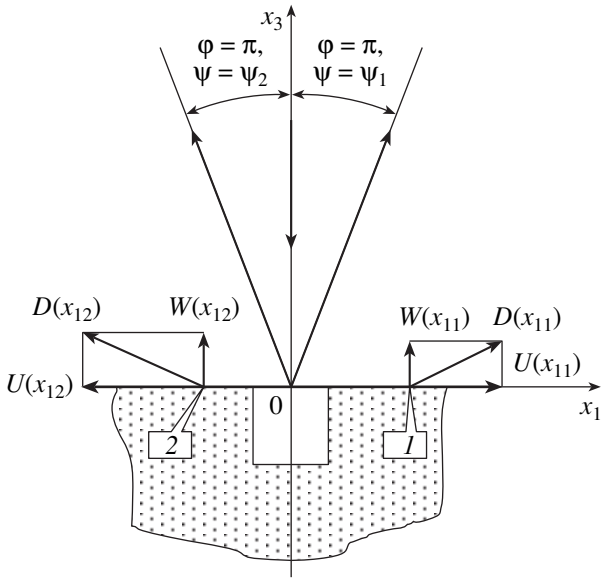


Fig. 1. Optical scheme of a holographic interferometer.

The rigid displacements, in turn, can be represented as the sum of translational and rotational components. It follows from Eqs. (4) and (5) that translations in the surface plane cancel out. Contributions from rotations about the axis normal to the surface are also absent, since their projections onto the coordinate axes at a small angle of rotation are independent of the coordinate x_1 and x_2. Translations that are normal to the surface only change the position of holographic interference fringes.

Finally, small rotations about the axis lying in the surface plane can be written as

$$W^0 = Ax_1 + Bx_2, \quad (7)$$

where A and B are unknown constants.

In view of (7), the differences between the displacements that are normal to the surface take the form

$$\begin{aligned}\Delta W_{1ij} &= \Delta w_{1ij} + A(x_{1i} - x_{1j}), \\ \Delta W_{2mn} &= \Delta w_{2mn} + B(x_{2m} - x_{2n}).\end{aligned} \quad (8)$$

For an arbitrary point on the surface with coordinates (x_1, x_2), the displacements $W(x_1, x_2)$ is found from the holographic interferogram obtained for $\psi \approx 0°$:

$$W(x_1, x_2) = \frac{\lambda N(x_1, x_2)}{1 + \cos\psi}. \quad (9)$$

With formulas (6), (8), and (9), we arrive at the relationship

$$\frac{\lambda N(x_1, x_2)}{1 + \cos\psi} - w(x_1, x_2) = Ax_1 + Bx_2. \quad (10)$$

In the case of a through hole of radius R drilled in a plate of thickness t, the quantity $w(x_1, x_2)$ depends on the difference in the principal stresses σ_1–σ_2 and in

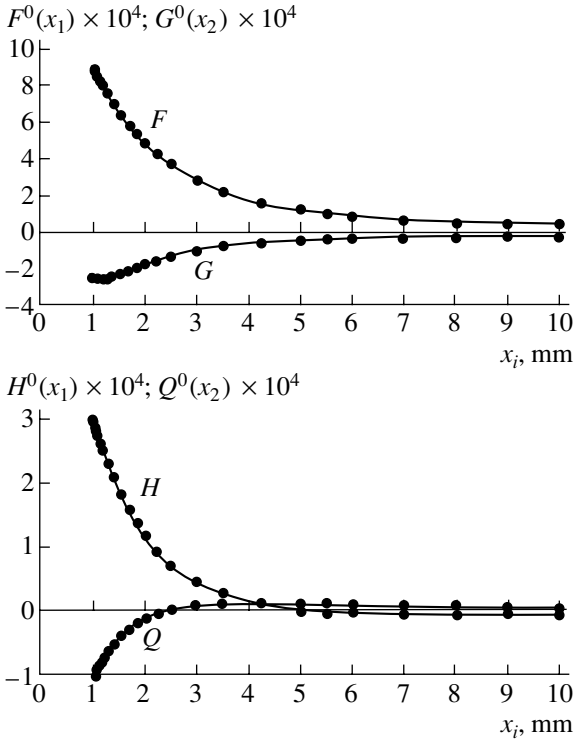


Fig. 2. Basis functions of unity displacement fields.

each point of the surface is given by [5]

$$w(x_1, x_2) = \frac{\nu t R^2 (x_1^2 - x_2^2)(\sigma_1 - \sigma_2)}{E(x_1^2 + x_2^2)^2}, \qquad (11)$$

where ν and E are the Poisson's ratio and the elastic modulus of the plate material.

For a blind hole, the value of $w(x_1, x_2)$ is found by numerically solving a related problem from the theory of elasticity. The set of Eqs. (10) and (11), which is based on calculated data, allows one to find the field of rotations and separate out the axes of principal residual stresses.

To raise the reliability of results, it is recommended to successively increase the diameter of the holes and record a hologram for each of the diameters. In this case, each of the holograms has its own field of rigid displacements, which makes it possible to determine the principal axis directions at small rotations.

In terms of the conventional linear elastic model, which is commonly used in such problems, the distributions of the differences in the strain components of the displacements $u(x_1)$, $u(x_2)$, $v(x_1)$, $v(x_2)$, $w(x_1)$, and $w(x_2)$ for pairs of points along the principal stress directions are written in the form

$$\begin{aligned} \Delta u_{ij} &= \sigma_1 \Delta F_{ij} + \sigma_2 \Delta G_{ij}, \\ \Delta w_{ij} &= \sigma_1 \Delta H_{ij} + \sigma_2 \Delta Q_{ij}, \\ \Delta v_{mn} &= \sigma_1 \Delta G_{mn} + \sigma_2 \Delta F_{mn}, \\ \Delta w_{mn} &= \sigma_1 \Delta Q_{mn} + \sigma_2 \Delta H_{mn}, \end{aligned} \qquad (12)$$

where ΔF, ΔG, ΔH, and ΔQ are the differences in calculated functions F, G, H, and Q of displacements caused by the unity stress acting sequentially along the directions x_1 and x_2.

For through holes, these functions are expressed analytically from solutions to a relevant problem in the theory of elasticity [5, 8].

For blind holes, these functions were obtained by the finite-element technique on a test object (a cube of edge a) placed in the uniform stress field $\sigma(x_1) = 1$. The drilling of holes of diameter $2R = a/10$ and depth $h_0 = 3R$ was simulated in the same way as in [9]. The basis distributions of the tangential longitudinal, $F^0(x_1)$, transverse, $G^0(x_2)$, and normal, $H^0(x_1)$ and $Q^0(x_2)$, displacements are given in Fig. 2. The functions F, G, H, and Q involved in Eq. (12) are found by merely multiplying the basis functions by the actual probing hole diameter $2R$ and taking into the oddness of F and G and evenness of H and Q.

With expressions (4), (5), (8), and (12), a set of equations for principal residual stresses can be written in the matrix form

$$\mathbf{Z} \cdot \mathbf{S} = \begin{bmatrix} Z_{11} & Z_{12} & \Delta x_1(1 + \cos\psi_k) & 0 \\ Z_{21} & Z_{22} & 0 & \Delta x_2(1 + \cos\psi_1) \end{bmatrix} \cdot \mathbf{S} = \lambda \mathbf{N}, \qquad (13)$$

where

$$\begin{aligned} Z_{11} &= \Delta F_{ij} \sin\psi_k + \Delta H_{ij}(1 + \cos\psi_k), \\ Z_{12} &= \Delta G_{ij} \sin\psi_k + \Delta Q_{ij}(1 + \cos\psi_k), \\ Z_{21} &= \Delta G_{mn} \sin\psi_1 + \Delta Q_{mn}(1 + \cos\psi_1), \\ Z_{22} &= \Delta F_{mn} \sin\psi_1 + \Delta H_{mn}(1 + \cos\psi_1), \end{aligned}$$

$\mathbf{S} = \{\sigma_1 \sigma_2 AB\}^T$ is the vector of desired quantities (the superscript T means transposition), and $\mathbf{N} = \{\Delta N_{1ij} \Delta N_{2mn}\}^T$ is the vector of fringe relative orders.

The solution of the set of Eqs. (13) by the least squares method yields (in matrix form)

$$\mathbf{S} = (\mathbf{Z}^T \cdot \mathbf{Z})^{-1} \cdot \mathbf{Z}^T \cdot \mathbf{N}. \qquad (14)$$

The sign of displacements and, hence, residual stresses is found from interferograms by analyzing the trajectories of a fringe reaching a probing hole [6].

The error in determining residual stresses is estimated with the computer simulation of experimental data. To this end, experimental data are statistically processed by the Monte Carlo method for a given inaccuracy of measuring the displacement vector components. It has been shown [6] that the tangential components of displacements are reliably and reproducibly determined at angles of observation between 40° and 60°.

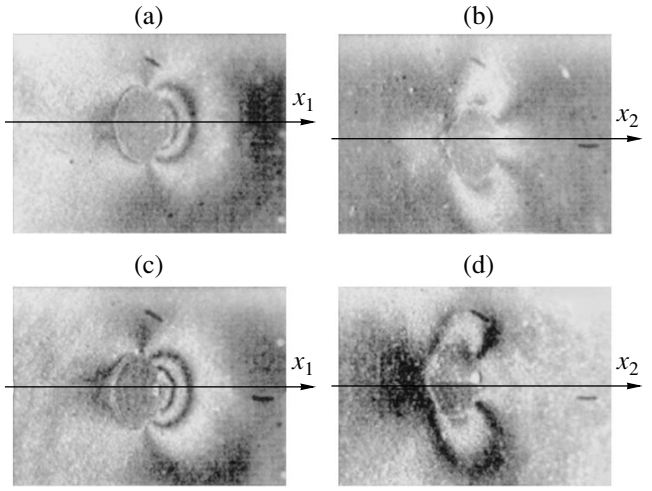


Fig. 3. Holographic interferograms of the surface near the probing hole in the steel beam strained by perfect bending. φ = (a) 0, (b) 90°, (c) 180°, and (d) 270°.

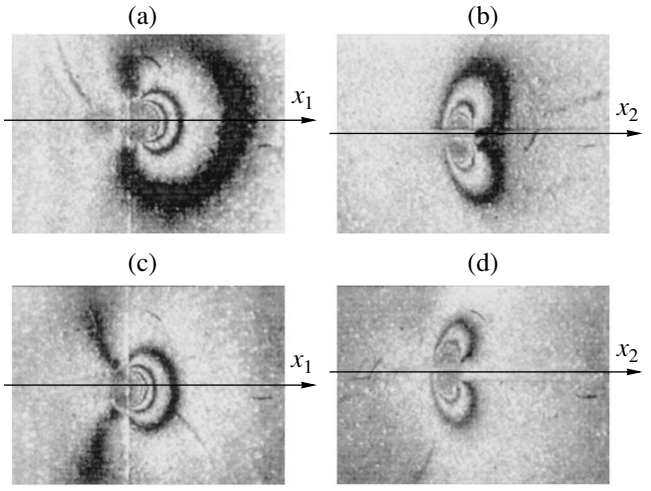


Fig. 4. Holographic interferograms of the thick-walled tube surface near the probing surface. (a–d) The same as in Fig. 1.

Our method was tested on an elastically strained 18 × 36 × 160-mm cantilever beam (its free end was subjected to a bending moment) made of St 40 steel. Under these conditions, the surface of the beam undergoes elastic uniaxial compression.

First, two holograms were taken when the surface was in the initial state. Either of two photoplates was mounted on a special device that makes it possible to remove the plates from the interferometer and return back to their position with high precision. Then, a hole 2 mm deep and 2 mm in diameter was drilled and one of the plates was exposed. Next, the diameter and depth of the hole were increased to 2.9 mm and the other plate was exposed.

Thus, two double-exposed holograms were recorded for two holes with various diameters. Figure 3 shows fringe patterns corresponding to the given stresses $\sigma_{x1} = -100$ MPa and $\sigma_{x2} = 0$. The angle of observation was $\psi = 50°$ in both cases. Even a cursory examination allows us to unambiguously separate out the principal axis. Fringes due to a rotation of the photoplate relative to the object are absent. The stresses found from overdetermined set (14) are $\sigma_{x1} = (-97 \pm 20)$ MPa and $\sigma_{x2} = (0.17 \pm 20)$ MPa; the unknown coefficients, $A = 3 \times 10^{-2}$ μm/mm and $B = -2.5 \times 10^{-3}$ μm/mm. The elastic stresses measured coincide with the given values within the computational error.

The method was also verified on a thick-walled flat-bottom pressurized tube. The tube was in the state of biaxial stress with a circumferential-to-axial stress ratio $\sigma_{x1} : \sigma_{x2} = 2 : 1$. Typical holographic interferograms of the surface near the probing hole for the given elastic stresses $\sigma_{x1} = 61$ MPa and $\sigma_{x2} = 30.5$ MPa are demonstrated in Fig. 4. Fringes due to a displacement or rotation of the plate relative to the object are absent. In the holograms, the axes of symmetry are aligned with the principal stress axes and the way the fringes reach the hole corresponds to biaxial extension. The stresses were measured to be $\sigma_{x1} = (60 \pm 12)$ MPa and $\sigma_{x2} = (31 \pm 12)$ MPa, which agrees well with the given values and ratio.

Our technique may be useful in practice, e.g., for determining residual stresses after welding. The holograms provide much information and allow one to find the direction, sign, and value of principal stresses.

ACKNOWLEDGMENTS

This work was supported by the Interindustry Research Center (project no. 774).

REFERENCES

1. *Handbook on Experimental Mechanics*, Ed. by A. Kobayashi (Prentice-Hall, Englewood Cliffs, 1989; Mir, Moscow, 1990), Vol. 2.
2. M. T. Flaman, Exp. Mech. **30**, 352 (1990).
3. A. A. Antonov, A. I. Bobrik, V. K. Morozov, *et al.*, Mekh. Tverd. Tela, No. 2, 41 (1980).
4. E. M. Beaney, Strain **12** (3), 99 (1976).
5. V. N. Bakulin and A. A. Rassokha, *Finite-Element Method and Holographic Interferometry in the Mechanics of Composites* (Mashinostroenie, Moscow, 1987).
6. V. S. Pisarev, V. P. Shchepinov, and A. Yu. Shchikanov, Zh. Tekh. Fiz. **66**, 99 (1996) [Tech. Phys. **41**, 50 (1996)].
7. Yu. I. Ostrovskiĭ, V. P. Shchepinov, and V. V. Yakovlev, *Holographic Interference Methods of Strain Measurement* (Nauka, Moscow, 1988).
8. D. G. Kurnosov and M. V. Yakutovich, Zavod. Lab. **12**, 960 (1946).
9. L. A. Antonov, A. I. Bobrik, V. K. Morozov, *et al.*, Mekh. Tverd. Tela, No. 2, 182 (1980).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1468–1474. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 111–117.

**OPTICS,
QUANTUM ELECTRONICS**

Modes of Optical Cavities Containing Selecting Components

V. N. Kudashov, A. B. Plachenov, and A. M. Radin

St. Petersburg State University of Low-Temperature and Food Technologies, St. Petersburg, Russia
e-mail: amradin@mail.ru

Received May 20, 2003

Abstract—A model of the field mode structure in optical cavities containing selecting components is suggested. The cavity also contains field-attenuating or field-enhancing media where elements of the 4×4 symplectic detour matrix of the cavity become complex. The model makes it possible to determine the stability of the cavity, non-Hermiticity of higher modes, and the complex astigmatism of the eigenmode field. Conditions for unilateral (unidirectional) and bilateral (bidirectional) stability are formulated. A cavity showing unilateral stability at the first transverse mode is described. © *2003 MAIK "Nauka/Interperiodica"*.

NOTATION AND DEFINITIONS

A four-dimensional complex column vector

$$Y = \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} \tag{1}$$

(**q** and **p** are two-dimensional column vectors) will be referred to as positive (negative) [1, 2] if such (positive or negative) is the quantity

$$\mathrm{Im}(\mathbf{p}^t\mathbf{q}^*), \tag{2}$$

where the superscript t means transposition and the asterisk, complex conjugation.

A subspace all nonzero vectors of which are positive (negative) will also be called positive (negative).

A 4×4 matrix T is called symplectic [3] if the equality

$$\sigma(TY_1, TY_2) = \sigma(Y_1, Y_2), \tag{3}$$

where

$$\sigma(Y_1, Y_2) = \mathbf{p}_1^t\mathbf{q}_2 - \mathbf{p}_2^t\mathbf{q}_1 \tag{4}$$

is the skew-symmetric product, is valid for any pair of vectors $Y_{1,2}$ of form (1).

Vectors whose components appear in the columns of a symplectic matrix form a symplectic basis.

Two vectors $Y_{1,2}$ are called skew-orthogonal if their skew-symmetric product (4) equals zero. A two-dimensional subspace all vectors of which are pairwise skew-orthogonal is called the Lagrangean plane [3].

PROPAGATION OF GAUSSIAN BEAMS IN FIRST-ORDER OPTICAL SYSTEMS

(1) Consider a first-order optical system [4, 5] where the Z axis is aligned with the optical axis of the system and the transverse coordinates are combined into a two-dimensional vector $\mathbf{r} = (x, y)^t$. Let, in a basic approximation, functions describing the state of the light field near the optical axis have the form

$$u^{(\pm)}(z, \mathbf{r}) = a(x, \mathbf{r})e^{\pm ik\tau(z, \mathbf{r})}, \tag{5}$$

where

$$\tau(z, \mathbf{r}) = \tau_0(z) + \frac{1}{2}\mathbf{r}^t H(z)\mathbf{r}, \tag{6}$$

and $H(z)$ is a 2×2 symmetric matrix. The time dependence is assumed to be harmonic.

In (5), signs "+" and "–" correspond to forward and backward waves propagating in the positive and negative Z directions, respectively, provided that $\tau_0(z)$ is an increasing function of z. It is assumed that the initial field-describing equations allow for the substitution of $-k$ for k, so that the functions $u^{(\pm)}$ simultaneously satisfy these equations with the same $\tau(z, \mathbf{r})$ and $a(z, \mathbf{r})$. The forward wave will have the form of a concentrated Gaussian beam if the matrix $H(z)$ has a positive definite imaginary part for any z. The backward wave is concentrated if the imaginary part of $H(z)$ is negative definite.

We also assume that the initial equations allow for solutions (in the form of (5) and (6)) such that the preexponential a in the basic approximation is **r**-independent, i.e., the dependence on the transverse coordinates is defined by the matrix H alone.

(2) The propagation of such beams is conventionally described in terms of *ABCD* ray matrices [6, 7]. Matrices $H_{\mathrm{in, out}} = H(z_{\mathrm{in, out}})$ that correspond to different values of the variable z are related as

$$H_{\mathrm{out}} = (C + DH_{\mathrm{in}})(A + BH_{\mathrm{in}})^{-1}, \tag{7}$$

where A, B, C, and D are 2×2 matrices that are the blocks of a 4×4 symplectic *ABCD* transformation

matrix

$$T = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

In the absence of field attenuation or field enhancement, the elements of the *A*, *B*, *C*, and *D* matrices are real. In optical systems containing selecting components (like Gaussian apertures), as well as field-attenuating (field-enhancing) media, the matrices become complex, with the *T* matrix remaining symplectic [8].

It is said that a beam has simple astigmatism if *H* is a diagonal matrix with different eigenvalues and complex astigmatism if *H* is nondiagonal. We *a fortiori* deal with complex astigmatism if the matrices *A*, *B*, *C*, and *D* are nondiagonal.

Field transformations when the matrix *H* in (5) and (6) is transformed according to (7) (*ABCD* transformation) may be generally represented (up to a factor) as a combination of elementary transformations: replacement of variables, multiplication by a Gaussian function, and Fourier transformation in one or both transverse coordinates. If the block *B* is a nondegenerate matrix, the representation may be in the form of integral operator [8]

$$(\mathbf{U}^{(\pm)}u)(\mathbf{r}) = \int_{\mathbb{R}^2} U^{(\pm)}(\mathbf{r}, \mathbf{r}')u(\mathbf{r}')d\mathbf{r}', \tag{8}$$

where the kernel is

$$U^{(\pm)}(\mathbf{r}, \mathbf{r}') = \frac{\pm k}{2\pi i\sqrt{\det B}} e^{\pm ik\Phi(\mathbf{r}, \mathbf{r}')},$$

$$\Phi(\mathbf{r}, \mathbf{r}') = \frac{1}{2}(\mathbf{r}'^t B^{-1}A\mathbf{r}' - 2\mathbf{r}'^t B^{-1}\mathbf{r} + \mathbf{r}^t DB^{-1}\mathbf{r}).$$

Note that the symplecticity of *T* means the symmetry of the matrices $B^{-1}A$ and DB^{-1}.

The product of *ABCD* matrices corresponds to a combination of operators *U*, which also is an integral operator of form (8) if the block *B* of the resultant matrix is nondegenerate.

(3) *ABCD* transformations are defined by the matrix *T* up to a factor, which does not depend on the form of the function. If the dependence of $u^{(\pm)}$ on the transverse coordinates is known for a given z_{in}, for an arbitrary z_{out} we have

$$u^{(\pm)}(z_{out}, \mathbf{r}) = \eta^{(\pm)}(z_{out}, z_{in})e^{\pm ik(\tau_0(z_{out}) - \tau_0(z_{in}))} \times \mathbf{U}^{(\pm)})(z_{out}, z_{in})u^{(\pm)}(z_{in}, \mathbf{r}). \tag{9}$$

Here, the operators $U^{(\pm)}(z_{out}, z_{in})$ correspond to *ABCD* matrices $T(z_{out}, z_{in})$ and are represented as combinations of the elementary transformations mentioned above. If the block *B* is nondegenerate, they can be represented as (8). Formula (9) involves, along with the eikonal factor $\exp\{\pm ik(\tau_0(z_{out}) - \tau_0(z_{in}))\}$, functions $\eta^{(\pm)}$ the specific form of which depends on the statement of the problem. These functions are by no means related to the transverse distribution form and can be calculated in the short wave approximation. The obvious properties of the functions $\eta^{(\pm)}$ are

$$\eta^{(\pm)}(z, z) = 1, \quad \eta^{(\pm)}(z_2, z_1)\eta^{(\pm)}(z_1, z_0) = \eta^{(\pm)}(z_2, z_0). \tag{10}$$

Later on, we assume that $\eta^{(\pm)}(z_{out}, z_{in})$ can be represented as

$$\eta^{(\pm)}(z_{out}, z_{in}) = \frac{\eta_1(z_{out})}{\eta_1(z_{in})}\eta_2^{(\pm)}(z_{out}, z_{in}), \tag{11}$$

with

$$\eta_2^{+}(z_{out}, z_{in}) = \eta_2^{(-)}(z_{in}, z_{out}). \tag{12}$$

The function η_1 describes the dependence of the field on local properties of the medium. By virtue of our assumption that the initial equations are symmetric with respect to the substitution of $-k$ for k, the function $\eta_1(z)$ does not depend on the wave propagation direction.

The function η_2 describes additional phase and/or amplitude steps (independent of the field transverse distribution) which arise when optical elements are located between z_{in} and z_{out}, for example, upon reflection from mirrors (the amplitude experiences a discontinuity if the mirror is not perfect). Relationship (11) means that waves propagating in opposite directions and passing through these optical elements suffer the same losses and exhibit the same phase shifts. In this respect, the behavior of the functions $\eta_2^{(\pm)}$ is akin to that of the eikonal factor $\exp\{\pm ik(\tau_0(z_{out}) - \tau_0(z_{in}))\}$.

MATRICES $H_{\pm}(z)$ IN THE CASE OF A BILATERALLY STABLE RING CAVITY

(1) In a ring optical cavity, the field *u* after complete detour transforms into itself:

$$u(z + l, \mathbf{r}) = u(z, \mathbf{r}) \tag{13}$$

(*l* is the total length of the cavity).

For functions like (5) and (6), this means, in particular, that

$$H(z + l) = H(z); \tag{14}$$

that is, if $z_{out} = z_{in} + l$, the matrix *H* is reproduced by transformation (7): $H_{out} = H_{in} = H$. Hence,

$$H = (C + DH)(A + BH)^{-1}, \tag{15}$$

where *A*, *B*, *C*, and *D* are the blocks of the detour matrix of the cavity (monodromy matrices [9]).

Monodromy matrices for different *z* are related to each other by similarity transformation. The presence of a solution $u^{(+)}$ or $u^{(-)}$ (see (5), (6)) concentrated at the

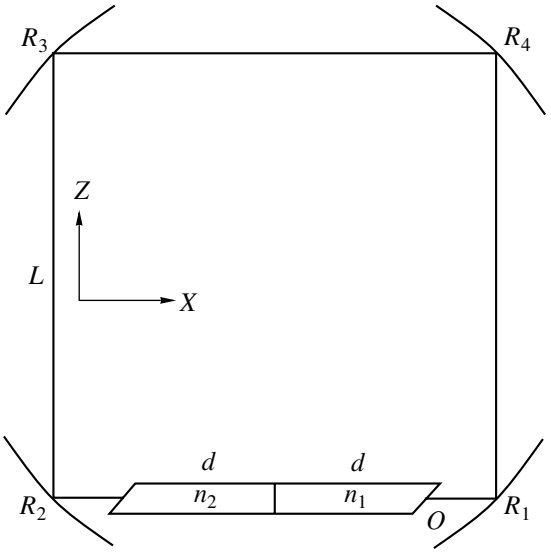


Fig. 1. Schematic of a ring cavity that has unilateral (unidirectional) stability at the wavelength $\lambda = 0.6328$ μm. The length of the cavity's arm is $L = 10$ cm. Mirrors R_1, R_2, R_3, and R_4 have radii of curvature of 34, 70, 100, and 200 cm, respectively. Absorbing apertures on the mirrors attenuate the field by a factor of e^{x^2/a^2}, where x is the transverse coordinate and $a = 10$ mm is the aperture width. Attenuating and enhancing media of length 4 cm each have refractive indices $n_1 = 1.01 + 0.05i$ and $n_2 = 1.01 - 0.05i$, respectively, and are located symmetrically about the center of the arm. The origin O is placed at the entrance to the absorbing medium. The number of the longitudinal mode is $N = 633375$.

axis of the cavity suggests, in particular, the presence of a matrix $H(z)$ that satisfies (14) and has the imaginary part that is positive or negative definite for any z. We will restrict our consideration to bilaterally stable cavities, which are of most practical interest because of the presence of concentrated solutions $u^{(\pm)}$ simultaneously satisfying (13). It should be noted that, if the monodromy matrix is not real, unilateral stability may occur. In this case, only one of oppositely propagating waves concentrates at the cavity axis [2], while the field of the other builds up with distance from the axis (instability). The physical reason for unilateral stability is field nonreciprocity. Earlier, field nonreciprocity was observed only in bilaterally stable cavities (see, e.g., [10]). However, the fact that it may cause unilateral stability has escaped the attention of the researchers. Figures 1 and 2 show a cavity that exhibits unilateral stability at the first transverse mode.

As follows from (9), the transverse distributions of the fields of the forward and backward waves over an arbitrary section z are eigenfunctions of the operators $\mathbf{U}^{(\pm)} = \mathbf{U}^{(\pm)}(z + l, z)$, which correspond to the monodromy matrix $T(z + l, z)$.

(2) As was shown [2], the bilateral stability of the cavity implies that the matrix T has two (positive and negative) invariant Lagrangean planes. The **q** and **p** components of vectors belonging to these planes are related by the relationships $p = H_{\pm}q$, where $H_{\pm}$ are desired symmetric solutions to Eq. (15) with imaginary parts of fixed sign. In this case, the matrix T allows for the representation as the product of three symplectic matrices:

$$T = W\Upsilon W^{-1}, \quad \Upsilon = \begin{pmatrix} M_- & 0 \\ 0 & M_+ \end{pmatrix}, \tag{16}$$

where $M_{\pm}$ are 2×2 matrices related as $M_-^{-1} = M_+^t$ because of symplecticity.

The first two, $Y_{1,2}^-$, and the last two, $Y_{1,2}^+$, columns of the matrix W belong to the negative and positive invariant Lagrangean planes of the matrix T, respectively. Let us represent W in block form:

$$W = \begin{pmatrix} Q_- & Q_+ \\ P_- & P_+ \end{pmatrix}. \tag{17}$$

Then, desired matrices $H_{\pm}$ can be written as

$$H_{\pm} = P_{\pm}Q_{\pm}^{-1}. \tag{18}$$

(3) If the matrix T is diagonalizable, $M_{\pm}$ may be taken in the form

$$M_{\pm} = \begin{pmatrix} e^{\pm i\theta_1} & 0 \\ 0 & e^{\pm i\theta_2} \end{pmatrix},$$

where $e^{\pm i\theta_{1,2}}$ are the eigenvalues of the matrix T and the columns $Y_{1,2}^+$ of the matrix W are the eigenvectors of T.

Generally, the quantities $\theta_{1,2}$ are complex. If $\theta_1 = -\theta_2$, invariant Lagrangean planes and, accordingly, the matrices $H_{\pm}$ are determined ambiguously: Eq. (15) has a continuous family of solutions [2, 11].

If the matrix T cannot be diagonalized,

$$M_+ = \begin{pmatrix} e^{i\theta} & e^{i\theta} \\ 0 & e^{i\theta} \end{pmatrix}, \quad M_- = \begin{pmatrix} e^{-i\theta} & 0 \\ -e^{-i\theta} & e^{-i\theta} \end{pmatrix},$$

provided that W is properly selected. In this case, the columns of W are both the eigenvectors and attached vectors of T.

(4) Let representation (16) apply to a monodromy matrix T_{in} that corresponds to a section $z = z_{\text{in}}$. Then, a matrix T_{out} for an arbitrary section $z = z_{\text{out}}$ has a similar representation with the same matrix Υ and the matrices $W_{\text{in, out}}$ are related to each other through the relationship

$$W_{\text{out}} = T(z_{\text{out}}, z_{\text{in}})W_{\text{in}},$$

where $T(z_{out}, z_{in})$ is *ABCD* transition from the plane $z = z_{in}$ to $z = z_{out}$.

Matrices

$$H_{\pm}(z_{in,\,out}) = P_{\pm}(z_{in,\,out})Q_{\pm}^{-1}(z_{in,\,out})$$

are related by equality (7). In a bilaterally stable cavity, the imaginary parts of $H_{\pm}(z)$ have definite sign at any z.

(5) If $z_{out} = z_{in} + l$, $T(z + l, z)$ is a monodromy matrix. Then, according to (16),

$$W(z + l) = W(z)\Upsilon,$$

and

$$Q_{\pm}(z + l) = Q_{\pm}(z)M_{\pm}, \quad P_{\pm}(z + l) = P_{\pm}(z)M_{\pm}.$$

Equality (14) will obviously be satisfied, since the common factor of the matrices $Q_{\pm}$ and $P_{\pm}$ does not affect the resultant matrix $H_{\pm}$ given by (18).

(6) Consider the functions

$$\psi_0^{\pm}(z, r) = \left(\frac{i}{2\pi}\right)^{1/2}\frac{k}{\sqrt{\det Q_{\pm}(z)}}e^{\pm ikr^t H_{\pm}(z)r/2}. \qquad (19)$$

For any fixed z, the functions ψ_0^+ and ψ_0^- satisfy the condition

$$\langle \psi_0^+(z, \mathbf{r}), \psi_0^-(z, \mathbf{r})\rangle = \int_{\mathbb{R}^2}\psi_0^+(z, \mathbf{r})\psi_0^-(z, \mathbf{r})d\mathbf{r} = 1, \qquad (20)$$

where $\langle , \rangle$ means the real scalar product over the transverse coordinates.

Also, at $z = z_{in,\,out}$, $\psi_0^{\pm}$, being the functions of the transverse coordinates, are related as

$$\psi^{(\pm)}(z_{out}, \mathbf{r}) = U^{(\pm)}(z_{out}, z_{in})\psi^{(\pm)}(z_{in}, \mathbf{r}).$$

After the complete detour of the cavity ($z_{out} = z_{in} + l$), the matrices $Q_{\pm}$ acquire factors $M_{\pm}$ and functions (19), factors $\lambda_0^{\pm} = (\det M_{\pm})^{-1/2}$. Clearly,

$$\lambda_0^{\pm} = e^{\mp i(\theta_1 + \theta_2)/2} \quad \text{or} \quad \lambda_0^{\pm} = e^{\mp i\theta}. \qquad (21)$$

Thus, for an arbitrary z, functions (19) are the eigenfunctions of the detour operators $U^{\pm}$ with eigenvalues (21).

From (9), it follows that waves counterpropagating at the fundamental mode are described by functions like

$$u_0^{\pm}(z, r) = \eta^{\pm}(z)e^{\pm ik\tau_0(z)}\psi_0^{\pm}(z, \mathbf{r}), \qquad (22)$$

where the function $\eta^{\pm}(z)$ equals the function $\eta^{\pm}(z, z')$ multiplied by a numerical factor for any z' (by virtue of (10), z' may take any value).

Possible values k determined from (13) will be given below.

BIRTH–DEATH OPERATORS

(1) Let us apply the operators $U^{(\pm)}$ to the vector functions $\pm\sqrt{i/k}\,\nabla u$ (∇ is the two-dimensional gradient along the transverse coordinates) and $\sqrt{k/i}\,\mathbf{r}u$. Differentiating (8) and then integrating by parts in view of the symplecticity of the matrix T, we arrive at the relationships

$$U^{(\pm)}(\pm\sqrt{i/k}\nabla u) = A^t(\pm\sqrt{i/k}\nabla U^{(\pm)}u) + C^t(\pm\sqrt{k/i}\mathbf{r}U^{(\pm)}u),$$

$$U^{(\pm)}(\sqrt{k/i}\mathbf{r}u) = B^t(\pm\sqrt{i/k}\nabla U^{(\pm)}u) + D^t(\sqrt{k/i}\mathbf{r}U^{(\pm)}u),$$

or, otherwise,

$$\mathbf{U}^{(\pm)}\begin{pmatrix}\pm\sqrt{i/k}\nabla\\ \sqrt{k/i}\mathbf{r}\end{pmatrix}u = T^t\begin{pmatrix}\pm\sqrt{i/k}\nabla\\ \sqrt{k/i}\mathbf{r}\end{pmatrix}U^{(\pm)}u. \qquad (23)$$

Note that commutation relationship (23) is general: it is valid for any operators $U^{(\pm)}$, including those represented otherwise than (8).

(2) Let Y be a vector of form (1). Consider the operators

$$\Lambda_Y^{(\pm)} = \pm\sqrt{i/k}(\mathbf{q}^t\nabla) + \sqrt{k/i}(\mathbf{p}^t\mathbf{r}).$$

These operators are conjugate in terms of real scalar product (20):

$$\langle\Lambda_Y^{(+)}u, v\rangle = \langle u, \Lambda_Y^{(-)}v\rangle. \qquad (24)$$

Performing left-sided multiplication of (23) by Y^t, we obtain

$$U^{(\pm)}\Lambda_Y^{(\pm)}u = \Lambda_{TY}^{(\pm)}U^{(\pm)}u. \qquad (25)$$

If $Y_{1,2} = \begin{pmatrix}\mathbf{q}_{1,2}\\ \mathbf{p}_{1,2}\end{pmatrix}$ are vectors and $\Lambda_{Y_{1,2}}^{(\pm)}$ are their associated operators, the commutator of these operators satisfies the equality

$$[\Lambda_{Y_1}^{(\pm)}, \Lambda_{Y_2}^{(\pm)}] = \pm\sigma(Y_1, Y_2). \qquad (26)$$

(3) Let $\{Y_1^-, Y_2^-, Y_1^+, Y_2^+\}$ be a symplectic basis. The operators

$$\Lambda_{\pm j} = \Lambda_{Y_j^{\pm}}^{(\pm)}, \quad \Lambda_{\pm j}^* = -\Lambda_{Y_j^{\mp}}^{(\pm)} \qquad (27)$$

satisfy the relationships

$$\text{(i) } [\Lambda_{\pm j}, \Lambda_{\pm i}] = [\Lambda_{\pm j}^*, \Lambda_{\pm i}^*] = 0,$$
$$\text{(ii) } [\Lambda_{\pm j}, \Lambda_{\pm i}^*] = \delta_{ij}, \qquad (28)$$
$$\text{(iii) } \Lambda_{\pm j}^t = \Lambda_{\mp j}^*,$$

which follow from (24), (26), and symplecticity condition (3). The operators $\Lambda^*_{\pm j}$ and $\Lambda_{\pm j}$ will be called, respectively, the birth and death operators.

HIGHER MODES

(1) Let $\{Y^{\pm}_{1,2}\}$ be the columns of the matrix W given by (17) and $\Lambda^*_{\pm j}$ and $\Lambda_{\pm j}$ be their associated birth and death operators (see (27)). Applying the birth and death operators to the functions $\psi^{\pm}_0$ (see (19)) yields the equalities

$$\Lambda_{\pm j}\psi^{\pm}_0(\mathbf{r}) = 0. \tag{29}$$

Let us introduce functions

$$\psi^{\pm}_{n_1 n_2} = \frac{1}{\sqrt{n_1! n_2!}}\Lambda^{*n_1}_{\pm 1}\Lambda^{*n_1}_{\pm 2}\psi^{\pm}_0. \tag{30}$$

They satisfy the real biorthogonality condition

$$\langle \psi^{+}_{n_1 n_2}, \psi^{-}_{m_1 m_2}\rangle = \delta_{n_1 m_1}\delta_{n_2 m_2}. \tag{31}$$

(2) Consider eigensubspaces of the operators $U^{(\pm)}$. First, we assume that the monodromy matrix T is diagonizable. Then, the vectors Y^{+}_j are eigenvectors for T and we obtain $U^{(\pm)}\Lambda^*_{\pm j} = e^{\mp i\theta_j}\Lambda^*_{\pm j}U^{(\pm)}$ from (25). From this equality and (30), it follows that

$$U^{(\pm)}\psi^{\pm}_{n_1 n_2} = \lambda^{(\pm)}_{n_1 n_2}\psi^{\pm}_{n_1 n_2},$$
$$\lambda^{(+)}_{n_1 n_2} = \exp\left\{\mp i\left[\left(n_1+\frac{1}{2}\right)\theta_1 + \left(n_2+\frac{1}{2}\right)\theta_2\right]\right\}. \tag{32}$$

Thus, the functions $\psi^{\pm}_{n_1 n_2}$ are eigenfunctions for the operators $U^{(\pm)}$ with the eigenvalues $\lambda^{(\pm)}_{n_1 n_2}$. Since the values of $\theta_{1,2}$ may be arbitrary (not only real) unlike the real case, eigenvalues (32) may not lie on the unit circle. Expression (32) implies the coincidence of the eigenvalues of the operators $U^{(+)}$ and $U^{(-)-1}$, which describe the transformation of the forward and backward waves after the detour of the cavity in directions coinciding with the directions of their propagation.

Multiple eigenvalues arise when either at least one of the values of $\theta_{1,2}$ is real and rationally comparable to π (or equals zero) or the arguments of $\theta_{1,2}$ coincide (or differ by π) and the moduli of $\theta_{1,2}$ are rationally comparable to each other.

(3) Let us elaborate on the case when the values of $\theta_{1,2}$ coincide up to sign, $\theta_1 = \pm\theta_2 = \theta$, and the matrix T has two two-dimensional eigensubspaces (at $\theta \neq 0, \pi$) corresponding to the eigenvalues $e^{\pm i\theta}$. The eigenvectors Y^{+}_j are then determined ambiguously, which, in turn, causes ambiguity in the determination of the operators $\Lambda^*_{\pm j}$ and $\Lambda_{\pm j}$, as well as functions (30) (and also functions (19) if $\theta_1 = -\theta_2$). Of course, such an ambiguity does not affect the structure of the eigensubspaces of the operators $U^{(\pm)}$.

(i) $\theta_1 = \theta_2 = \theta$. Subspaces $X^N_{\pm}$, the linear shells of the functions $\{\psi^{+}_{n_1 n_2},\ n_1 + n_2 = N\}$ (N is nonnegative), are the $(N+1)$-dimensional eigensubspaces of the operators $U^{(\pm)}$ with the eigenvalues $\exp\{\mp i(N+1)\theta\}$.

(ii) $\theta_1 = -\theta_2 = \theta$. Subspaces $\tilde{X}^N_{\pm}$, the linear shells of the functions $\{\psi^{(\pm)}_{n_1 n_2},\ n_1 - n_2 = N\}$ (N is arbitrary), are the infinitely dimensional eigensubspaces of the operators $U^{(\pm)}$ with the eigenvalues $\exp\{\mp i N\theta\}$.

With the basis vectors $Y^{\pm}_j$ taken otherwise, the functions $\{\psi^{(\pm)}_{n_1 n_2}\}$ change but belong to the same invariant subspaces. In particular, the continuous families of the functions $\psi^{\pm}_0$ (see (19)) belong to the subspaces $\tilde{X}^0_{\pm}$ if $\theta_1 = -\theta_2$.

An additional expression arises if θ is real and rationally comparable to π.

(4) Let the matrix T be nondiagonizable. In this case, the functions ψ^{+}_{0n} are the eigenfunctions of the operator $U^{(+)}$. It is easy to find that

$$U^{(+)}\psi^{+}_{n,N-n} = e^{-i(N+1)\theta}\sum_{j=0}^{n}\frac{(-1)^{n-j}}{(n-j)!}\sqrt{\frac{n!(N-j)!}{j!(N-n)!}}\psi^{+}_{j,N-j}. \tag{33}$$

Let us designate the linear shell of the functions $\{\psi^{+}_{0,N}, \psi^{+}_{1,N-1}, \ldots, \psi^{+}_{N,0}\}$ as X^N_{+}. From (33), it follows that the subspace X^N_{+} is invariant for the operator $U^{(+)}$ and the matrix of contraction into this subspace, $U^{(+)}$, in the basis selected has the upper triangular shape.

The functions ψ^{-}_{n0} are the eigenfunctions of the operator $U^{(-)}$. In addition,

$$U^{(-)}\psi^{-}_{n,N-n} = e^{i(N+1)\theta}\sum_{j=n}^{N}\frac{1}{(j-n)!}\sqrt{\frac{j!(N-n)!}{n!(N-j)!}}\psi_{j,N-j}. \tag{34}$$

Similarly, designating the linear shell of the functions $\{\psi^{-}_{0,N}, \psi^{-}_{1,N-1}, \ldots, \psi^{-}_{N,0}\}$ as X^N_{-}, we see that the subspace X^N_{-} is invariant for the operator $U^{(-)}$ and the matrix of contraction into this subspace, $U^{(-)}$, in the basis selected has the lower triangular shape.

EIGENFUNCTIONS OF THE CAVITY AND EIGENVALUES OF WAVENUMBERS

(1) Above, we constructed a set of eigenfunctions and eigenvalues of the operators $U^{(\pm)}$, which describe (up to a factor) the transformation of the forward and backward waves after the detour of the cavity in the positive Z direction. Now we may proceed to a construction of solutions satisfying condition (13). According to (9), the transverse field distribution in a given section is bound to satisfy the condition

$$u^{(\pm)}(\mathbf{r}) = c_{\pm}e^{\pm ik\Delta\tau}(U^{(\pm)}u^{(\pm)})(\mathbf{r}),$$

from which one can find, in particular, the eigenvalues of the wavenumbers k of the cavity. Here, $\Delta t = \tau_0(z + l) - \tau_0(z)$ and $c_{\pm} = \eta^{\pm}(z + l, z) = \eta_2^{\pm}(z + l, z)$ (the functions η_1 do not contribute to $c_{\pm}$ by virtue of (11)). Note that $c_- = c_+^{-1}$, as follows from (12).

The functions $u^{(\pm)}$ are the eigenfunctions of the operators $U^{(\pm)}$; that is, they coincide with functions (30) (or their linear combinations if the eigenvalues coincide) up to a factor. Then, if the matrix T is diagonizable,

$$c_{\pm}e^{\pm ik\Delta\tau}\lambda_{mn}^{(\pm)} = 1. \quad (35)$$

Since $c_- = c_+^{-1}$ and $\lambda_{mn}^{(-)} = \lambda_{mn}^{(+)-1}$, expressions (35) specify the same set of conditions, which, in view of (32), can be represented in the form

$$-i\ln c + k\Delta\tau - \left[\left(m + \frac{1}{2}\right)\theta_1 + \left(n + \frac{1}{2}\right)\theta_2\right] = 2\pi N,$$

where $c = c_+$ and N is a natural number.

Now we can determine a discrete set of k:

$$k_{Nmn} = \left\{2\pi N + \left[\left(m + \frac{1}{2}\right)\theta_1 + \left(n + \frac{1}{2}\right)\theta_2\right] + i\ln c\right\}/\Delta\tau. \quad (36)$$

If T is nondiagonizable,

$$k_{Nm} = \{2\pi N + (2m + 1)\theta + i\ln c\}/\Delta\tau. \quad (37)$$

(2) Let $\{\psi_{nm}^{\pm}(z_0, r)\}$ be a set of functions (30) that are constructed in the section $z = z_0$. Using the operators $U^{(\pm)}(z, z_0)$, we extend them throughout the cavity:

$$\psi_{nm}^{(\pm)}(z, r) = U^{(\pm)}(z, z_0)\psi_{nm}^{(\pm)}(z_0, r). \quad (38)$$

For any section z, functions (38) are represented in the form of (30), where the operators $\Lambda_{\pm j}^{*}(z)$ are given by the columns of the matrix $W(z)$. The biorthogonality condition over the transverse coordinates (see (31)) also remains valid:

$$\langle\psi_{nm}^{+}(z, r), \psi_{st}^{-}(z, r)\rangle = \delta_{ns}\delta_{mt}.$$

(3) In view of the aforesaid, we finally obtain a set of the cavity's eigenfunctions (that correspond to k_{Nnm}):

$$u_{nm}^{\pm}(z, r) = \eta^{\pm}(z)\exp(\pm ik_{Nnm}\tau_0(z))\psi_{nm}^{\pm}(z, r),$$

where the functions $\eta^{\pm}(z)$ is defined in the same way as in (22).

If T is nondiagonizable, the eigenfunctions corresponding to k_{Nn} are

$$u_n^{+}(z, r) = \eta^{+}(z)\exp(ik_{Nn}\tau_0(z))\psi_{0n}^{+}(z, r),$$
$$u_n^{-}(z, r) = \eta^{-}(z)\exp(-ik_{Nn}\tau_0(z))\psi_{n0}^{-}(z, r). \quad (39)$$

(4) Generally, the values of k given by (36) and (37) are complex. Then, strictly speaking, a fixed sign of the imaginary part of the matrix kH, rather than H, should be taken as the condition that the solutions $u^{(\pm)}$ are concentrated. Furthemore, the fact that k is not real will affect the time dependence of the wave field: the harmonic component will be imposed on the exponential growth (if the field is enhanced) or decay (if the field is attenuated). For the growth (decay) not to be too fast, the imaginary part of k (at sufficiently large N) must be much smaller than the real part. Then, the complexity of k will have a minor effect on the concentration of the

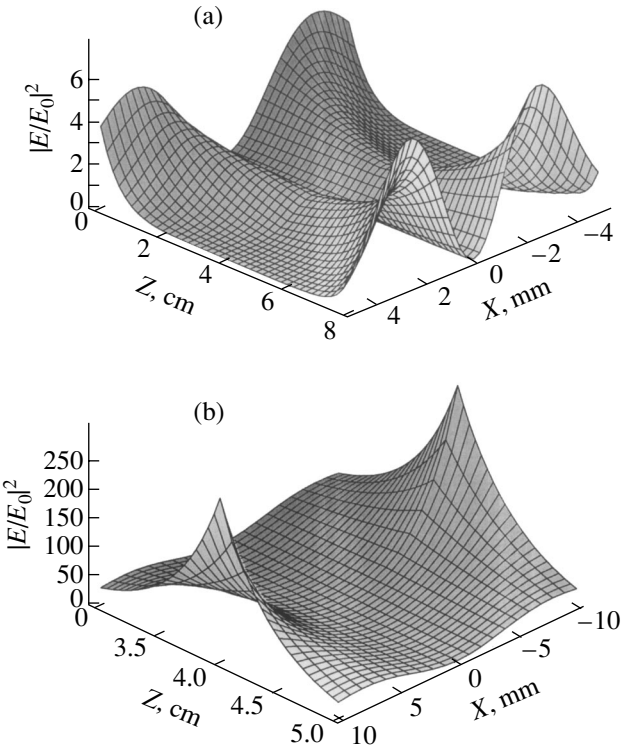


Fig. 2. Transverse distribution of the relative intensity of the first transverse mode for waves propagating in the (a) positive and (b) negative Z direction. In panel (a), the wave is stable in the active region of the cavity and, hence, throughout the cavity. In panel (b), the wave becomes locally unstable over the length between 3.7 and 4.3 cm. Since the counter wave (panel (a)) is stable throughout the cavity, the cavity is unilaterally (unidirectionally) stable (panel (b)).

solution (the imaginary parts of H and kH will have fixed sign simultaneously).

CONCLUSIONS

Thus, we (i) represented higher modes of counterpropagating waves in terms of birth–death operators and established a relationship between the birth and death operators for these waves, (ii) established the biorthogonality of the mode sets for counterpropagating waves in terms of real scalar product, (iii) found the eigenvalues of the wavenumbers of ring optical cavities in terms of the eigenvalues of *ABCD* matrices, and (iv) gave an example of a cavity that exhibits unilateral stability at the first transverse mode (Figs. 1, 2).

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education of the Russian Federation (grant no. E00-3.2-164).

REFERENCES

1. Yu. L. Daletskiĭ and M. G. Kreĭn, *Stability of Solutions to Differential Equations in Banach Space* (Nauka, Moscow, 1970).
2. V. N. Kudashov, A. B. Plachenov, and A. M. Radin, Opt. Spektrosk. **93**, 851 (2002) [Opt. Spectrosc. **93**, 781 (2002)].
3. V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Nauka, Moscow, 1989; Springer Verlag, New York, 1989).
4. M. J. Bastiaans, Optik (Stuttgart) **88** (4), 163 (1991).
5. I. V. Golovnin, A. I. Kovrigin, A. N. Konovalov, and G. D. Laptev, Kvantovaya Élektron. (Moscow) **22**, 461 (1995).
6. H. Kogelnik, Bell Syst. Tech. J. **44**, 455 (1965).
7. A. Gerrard and J. M. Burch, *Introduction to Matrix Methods in Optics* (Wiley, New York, 1975; Mir, Moscow, 1978).
8. V. N. Kudashov, A. B. Plachenov, and A. M. Radin, Opt. Spektrosk. **88**, 330 (2000) [Opt. Spectrosc. **88**, 291 (2000)].
9. V. M. Babich and V. S. Buldyrev, *Asymptotic Methods in Problems of Short Wave Diffraction* (Nauka, Moscow, 1972).
10. Yu. A. Anan'ev, *Optical Cavities and Laser Beams* (Nauka, Moscow, 1990).
11. V. N. Kudashov, A. B. Plachenov, and A. M. Radin, Opt. Spektrosk. **88**, 130 (2000) [Opt. Spectrosc. **88**, 121 (2000)].

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1475–1479. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 118–122.

ACOUSTIC, ACOUSTOELECTRONICS

Acoustooptic Cells with Nonuniform Length of Light–Sound Interaction

V. B. Voloshinov and G. A. Knyazev

Moscow State University, Vorob'evy gory, Moscow, 119992 Russia

e-mail: volosh@osc162.phys.msu.su

Received February 7, 2003; in final form, May 14, 2003

Abstract—Diffraction of light by acoustic waves that are generated in the acoustooptic cells of piezoelectric transducers with complex geometry is studied. The diffraction by acoustic beams with triangular, quadrangular (rhombic), hexagonal, etc., cross sections, when the lengths of light–sound interaction in the cross-sectional area of the light beam are different, is considered in the plane wave approximation. The difference in the length of interaction affects the instrument function of acoustooptic devices and provides the suppression of the side lobes in their transmission function. The advantages of using cells with complex-geometry transducers in tunable acoustooptic filters that are incorporated into fiber-optic communication lines with wavelength-division multiplexing (WDM) are discussed. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

Acoustooptic interaction in glasses and crystals is of both scientific and applied interest [1–5]. Diffraction of light by ultrasound attracts special attention, because light–ultrasound interaction is widely used for controlling optical beams. Acoustooptic modulators, deflectors, and filters, as well as optical data processing devices, have found a wide utility in optics and spectroscopy, laser physics and technology, and data transmission and processing systems [3–10].

An acoustooptic cell is a glass or crystalline unit where piezoelectric transducers generate acoustic waves [1, 2]. Usually, the transducers have the form of thin rectangular plates. The size of the light–sound interaction region in the cell depends on the transverse size of an acoustic column (beam). The cross section of an acoustic column is a rectangle of length l and width d, both nearly coinciding with the sizes of a driving (converting) electrode. As a rule, the width d of the column exceeds the light beam diameter to provide the maximal efficiency of diffraction over the cross section of the light beam, suppress edge effects, and decrease the divergence of ultrasound. As for the acoustic column length l, it is constant for all partial light rays within the cross section of the light beam in the light propagation direction [1–3].

If the ultrasound power is uniformly distributed over the transducer's aperture, the general directivity of acoustic energy is described by the function $z = \sin c^2\xi$ [4–6]. This function has a main lobe and symmetrically arranged side lobes with decreasing intensity. The side directivity of acoustic energy depends on the transducer length l. The transfer function of an acoustooptic cell, i.e., the dependence of the diffracted intensity I_d on the angle of incidence θ and optical wavelength λ (or ultrasonic frequency f), also has a main maximum and side maxima [4, 6–9].

If the diffraction efficiency is maximal, $I_d/I_i = 1.0$, where I_i is the intensity of incident light, the optical energy in side lobes may be significant. For example, with $I_d/I_i = 1.0$, the intensity escaped into side maxima with the numbers ± 1 exceeds 10%. The presence of side lobes in the transfer function is an obvious disadvantage of acoustooptic devices [7–10]. In acoustooptic signal-selecting filters used in fiber-optic communication lines, the presence of side maxima adversely affects the spectral characteristics of the device and rises the crosstalk level [10].

CONTROL OF THE INSTRUMENT FUNCTION OF ACOUSTOOPTIC CELLS

To date, several ways of reducing the intensity of side lobes in the instrument function of acoustooptic cells have been devised [3–5]. Side lobes may be suppressed, i.e., by applying sectional transducers with apodization [4, 5, 7]. Transducers with unequal lengths of sections are also efficient in this respect [7]. However, such ways of controlling the transmission (instrument) function of acoustooptic cells appear to be very complex. A simpler approach is the use of piezoelectric transducers with driving electrodes having the shape of a rhombus, hexagon, trapezoid, ellipse, etc. in the cross section [8, 9]. In this work, side lobes are suppressed by using acoustooptic filter cells that have a nonuniform length of acoustooptic interaction over the cross section of the light beam [3–5]. If the acoustooptic interaction length l is unequal for partial rays in an incident light beam, conditions for the side lobe appearance for each

1063-7842/03/4811-1475$24.00

of the rays will also differ. Therefore, the position of the side lobes of the cell instrument function relative to the main (central) lobe will depend on the length of interaction. Thus, the transmission function at the exit from such devices will have another shape compared with that in the case of a conventional rectangular electrode.

Unlike known examples of applying complex-geometry transducers in acoustooptic devices [4, 8, 9], we consider the case where the maximal length l and width d of the transducer are fixed. It is assumed that the shape of the driving electrode is varied within the sizes of a rectangular transducer. In practice, this restriction arises, i.e., when the length l of a piezoelectric plate is limited by the size of an acoustooptic crystal, its optical homogeneity, or light absorption conditions and the width of the plate is limited by the light beam aperture (that is, $d \geq a$).

INTENSITY OF DIFFRACTED LIGHT

In an acoustic cell with a rectangular transducer (Fig. 1a), the diffracted intensity varies with incident light intensity by the law

$$I_d = I_i \frac{q^2}{q^2 + \eta^2} \sin^2\left[\sqrt{q^2 + \eta^2}\frac{l}{2}\right], \quad (1)$$

where η is the mismatch parameter and q is the acoustooptic coupling coefficient, which depends on the elastic strain, photoelastic properties of the medium, optical wavelength, Bragg angle, acoustooptic figure of merit M_2, acoustic power P, and cross-sectional area $S = ld$ of an acoustic column [1–4].

With the phase matching condition satisfied, η identically equals zero. It can be shown that the diffraction intensity in this case is given by

$$I_d = I_i \sin^2\left[\frac{\pi l}{\lambda \cos\theta}\sqrt{\frac{M_2 P}{2S}}\right]. \quad (2)$$

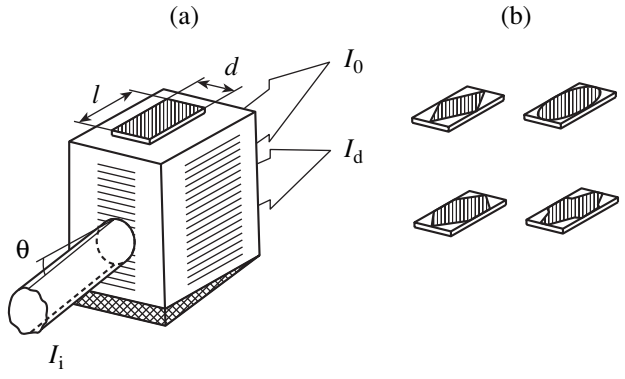


Fig. 1. Acoustooptic cell and piezoelectric transducers with (a) rectangular and (b) other possible configurations of driving electrodes.

From (1) and (2), it follows that, if the control power $P = P_0$, where $P_0 = 0.5S\lambda^2\cos^2\theta/M_2 l^2$, the diffracted light intensity is maximal. The product of the acoustooptic coupling coefficient and the length of interaction is $ql = \pi$ in this case. Clearly, the condition $ql = \pi$ is valid for all partial rays in the cross section of the optical beam shown in Fig. 1a, since the length of interaction for them is the same.

DIFFRACTION EFFICIENCY FOR A NONUNIFORM LENGTH OF INTERACTION

When the cross section of an acoustic column generated by the piezoelectric transducer is other than rectangular, the diffraction pattern is different. Some of the driving electrode configurations in the transducers are shown in Fig. 1b (note once again that the maximal sizes of driving electrodes are limited by the sizes of a rectangular transducer). In experiments, we studied electrodes of different configurations; here, results on the diffraction efficiency are given for only triangular, rhombic, and hexagonal electrodes. These results are easy to generalize for transducers of more intricate shape, e.g., elliptic and Gaussian.

A piezoelectric transducer with a rhombic electrode is shown in Fig. 2a. Here, the length of acoustooptic interaction is seen to depend on the coordinate x when the light beam is directed along the y axis. If the driving electrode is of triangular shape (Fig. 2b), the interaction length is given by $l(x) = l_0 - |x|\cot\alpha$, where $|x| \leq 0.5d$. In this formula, the angle α defines the shape of the driving electrode. This formula is also valid for transducers in the form of a trapezoid, rhombus, hexagon, etc., which immediately follows from the symmetry of the problem considered. Thus, in view of (1), the diffracted intensity for each partial optical ray within the acoustic column aperture is given by

$$I_d = I_i \frac{q^2}{q^2 + \eta^2} \sin^2\left[\sqrt{q^2 + \eta^2}\frac{l_0 - |x|\cot\alpha}{2}\right]. \quad (3)$$

With the phase matching condition satisfied, relationship (3) yields the diffracted intensity for each partial ray in the light beam cross section. The diffraction resulting efficiency $\bar{I}_d$ at the exit from the cell is the sum of the intensities of all diffracted rays in the light beam cross section of diameter $a = d$:

$$\bar{I}_d = \frac{1}{d}\int_{-d/2}^{d/2} I_d(x)dx = \frac{2}{d}\int_0^{d/2} I_i \sin^2\left[\frac{q}{2}(l_0 - |x|\cot\alpha)\right]dx. \quad (4)$$

As follows from (4), a cell with a nonuniform length of acoustooptic interaction basically cannot provide the maximal diffraction efficiency $\bar{I}_d = 1.0$. This is because, with the coupling factor q fixed, the product $ql(x)$ equals π only if $|x| = (l_0 - \pi/q)\tan\alpha$. For all other

values of x, this product is other than π. Therefore, the total diffracted intensity is less than maximal.

Our study also supports the fundamental inference that the diffraction resulting efficiency $\bar{I}_d$ depends on the driving electrode shape for a given ultrasound power P. As follows from (4), the diffraction efficiency is a function of the angle α. Figure 3 shows a family of total diffracted intensity vs. normalized ultrasound power (P/P_0) curves for various α. The diffraction efficiency is maximal, $\bar{I}_d = 1.0$, only in the case of a rectangular electrode. As α decreases, so does the diffraction efficiency.

Figure 3 suggests that cells with nonrectangular transducers provide the maximal values of $\bar{I}_d$ if the ultrasound power P exceeds the power P_0 that is typical of acoustic waves generated by a rectangular transducer. Calculations show that a decrease in the angle α causes the acoustic power to grow compared with the conventional (rectangular) case. For example, at $\alpha = \alpha_4$ (Fig. 3, curve *4*), the resulting intensity turns out to be smaller than unity: $\bar{I}_d = 0.85$. The control signal power in this case rises to $P = 1.6P_0$.

Thus, if the transducer in the cell is nonrectangular, the maximal diffraction efficiency somewhat drops, while the control acoustic power increases.

To conclude, it should be noted that the limits of integration in (5) may be changed if the light beam diameter is smaller than the transducer's width. Then, relationship (5) also applies to the analysis of diffraction efficiency in the case of transducers with pentagonal, hexagonal, etc., electrodes. If the driving electrode shape is such that the interaction length vs. coordinate is described by the elliptical, Gaussian, or any other dependence, appropriate substitutions in expression (4) make it possible to estimate the diffraction parameters in these cases too.

DIFFRACTION EFFICIENCY UNDER THE PHASE MISMATCH CONDITION

The dependence of the diffracted intensity on the angle of incidence on ultrasound and ultrasound frequency (or optical wavelength) defines the form of the transfer function of an acoustooptic cell. The transfer function and the intensity of side lobes can be found by analyzing acoustooptic interaction when the phase matching condition is violated. Formula (3) yields the following expression for the total diffraction efficiency $\bar{I}_d$ at the exit from an acoustooptic cell:

$$\bar{I}_d = I_i \frac{q^2}{2(q^2+\eta^2)} \times \left\{1 - \frac{\sin(l_0\sqrt{q^2+\eta^2}) - \sin[(l_0 - d\cot\alpha)\sqrt{q^2+\eta^2}]}{d\cot\alpha\sqrt{q^2+\eta^2}}\right\}. \quad (5)$$

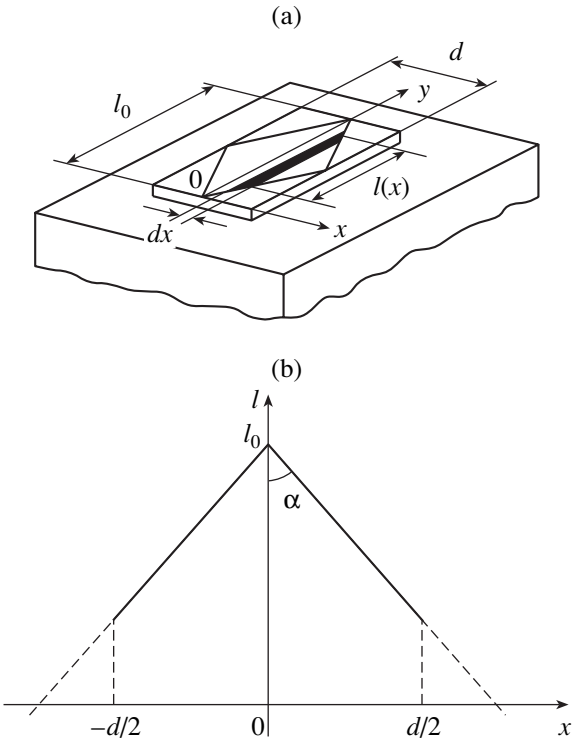


Fig. 2. (a) Estimation of the effective interaction length for transducers with the rhombic shape of the electrode. (b) Pentagonal and triangular shapes of the electrode.

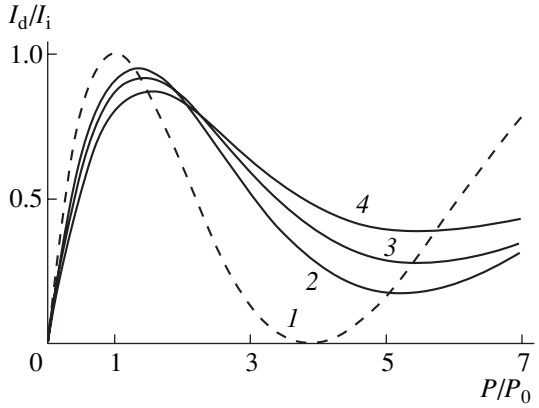


Fig. 3. Diffraction efficiency vs. normalized acoustic power. α = (*1*) 90°, (*2*) 75°, (*3*) 60°, and (*4*) 45°.

The diffraction efficiency vs. mismatch parameter for various powers and angles α is presented in Fig. 4. The calculations were performed for the maximal interaction length 1.0 cm. At small α, the transmission function of an acoustooptic cell becomes smooth, i.e., does not exhibit oscillations. Side lobes smooth out to the greatest extent, minimizing the possibility of signal escape through one of the side lobes of the instrument function.

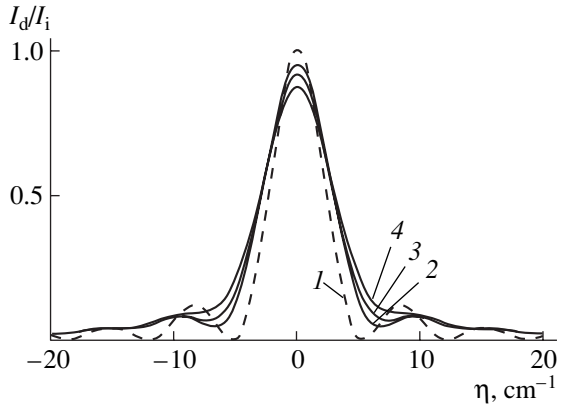


Fig. 4. Diffraction efficiency vs. mismatch parameter for transducers with different configurations. (*1*) $\alpha = 90°$, $P = P_0$; (*2*) $\alpha = 75°$, $P = 1.3P_0$; (*3*) $\alpha = 60°$, $P = 1.4P_0$; and (*4*) $\alpha = 45°$, $P = 1.5P_0$.

Earlier, it was shown [9] that the smoothing out of side lobes is coincident with the broadening of the main lobe by a factor of no more than 1.5–2.0 compared with the case of a rectangular transducer. This is also true for electrodes whose shape is described by a Gaussian, sinusoidal, or step function.

As follows from Fig. 4, a change in the driving electrode shape suppresses noticeably the side lobes of the transfer function of a cell. With the electrode shape properly selected, the intensity of each of the side lobes may be halved compared with that for a rectangular transducer. Side lobes may be suppressed still further if no restrictions are imposed on the electrode maximal dimensions. To this end, one should decrease the width d of the piezoelectric transducer keeping the piezoelectric plate length l_0 fixed or, conversely, increase the length of the plate for a given width d [8, 9].

CROSSTALK SUPPRESSION AT THE EXIT FROM FILTERS

Tunable optical filters intended for the spectral interval $\Delta\lambda = 1532–1565$ nm are promising acoustooptic devices for WDM fiber-optic communication lines [10]. The wavelength spacing $\delta\lambda$ between neighboring channels and the total number of channels $N = \Delta\lambda/\delta\lambda$ in a fiber-optic communication line depend on the crosstalk level at the exit from an acoustooptic device. It is known that the crosstalk level in communication lines must not exceed –20 dB. This means that, if $\eta = 0$ and an acoustooptic filter is tuned to transmit radiation with a wavelength λ_0 through one of the channels, the stray transmission of the filter at a wavelength $\lambda_1 = \lambda_0 \pm \delta\lambda$ must be less than 1%; that is, $I_d/I_i \le 0.01$ [10].

For an acoustooptic filter incorporated into a WDM communication line, it is important that the spectrum of light incident on the device be discrete, since all optical signals in the fiber are monochromatic. Therefore, the crosstalk level at the exit from the filter does not directly depend on the width of its passband: it is defined by a value of the transmission function at the wavelength of the neighboring channel. For the transducer with the rectangular electrode length $l = 1.0$ cm, calculations by formula (1) yield a diffraction efficiency $I_d/I_i \le 0.01$ at a mismatch parameter $|\eta| \ge 35$ cm^{-1}. If the driving electrode of the transducer has the shape of a triangle, rhombus, or hexagon, then, as follows from formula (6), a filter with $l_0 = 1.0$ cm has a transmission coefficient of less than 1% at a mismatch parameter $\eta \ge 20$ cm^{-1}. Thus, for the same stray signal level (1%), the mismatch parameter in the filter with the complex-shape electrode is roughly half as large as in the device with the rectangular transducer because oscillations in the spectral characteristic are absent.

It is known that the mismatch due to acoustooptic interaction equals the optical wavelength difference $\delta\lambda = \lambda_0 - \lambda_1$ [3, 4]. Therefore, the less the spectral spacing $\delta\lambda$ between neighboring channels, the lower the mismatch. Thus, the use of complex-shape transducers in the filters makes it possible to decrease the spectral spacing of optical carriers and, thereby, raise the total number of signals in fiber-optic communication lines. Therefore, the performance of acoustooptic filters with complex-shape transducers is superior to that of conventional filters with rectangular transducers.

CONCLUSIONS

In this work, we studied acoustooptic cells with nonrectangular electrodes of piezoelectric transducers. The use of complex-geometry transducers allows one to suppress side lobes in the transmission function of the cells. Accordingly, the crosstalk level at the exit from acoustooptic filters incorporated into WDM fiber-optic communication lines may be substantially reduced compared with filters using transducers of conventional rectangular geometry.

The improvement of the device performance by using complex-geometry transducers is attendant with the broadening of the passband and a decrease in the diffracted intensity. However, the inevitable degradation of these parameters is insignificant and virtually does not affect filtration parameters in WDM fiber-optic communication lines. Therefore, the way of improving the acoustooptic device performance that is presented in this work may be applied in fiber-optic technology owing to its simplicity and versatility.

REFERENCES

1. V. I. Balakshy, V. N. Parygin, and L. E. Chirkov, *Physical Principles of Acousto-Optics* (Radio i Svyaz', Moscow, 1985).

2. A. Korpel, *Acousto-Optics* (Marcel Dekker, New York, 1988; Mir, Moscow, 1993).
3. J. Xu and R. Stroud, *Acousto-Optic Devices* (Wiley, New York, 1992).
4. A. Goutzoulis and D. Pape, *Design and Fabrication of Acousto-Optic Devices* (Marcel Dekker, New York, 1994).
5. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, New York, 1968; Mir, Moscow, 1970).
6. I. C. Chang, Proc. SPIE **90**, 12 (1976).
7. V. N. Parygin, A. V. Vershoubsky, and E. Yu. Filatova, Zh. Tekh. Fiz. **71** (9), 73 (2000) [Tech. Phys. **45**, 1045 (2000)].
8. L. Bademian, Opt. Eng. **25**, 303 (1986).
9. D. Pape, P. Wasilouski, and M. Krainak, Proc. SPIE **789**, 116 (1987).
10. J. Sapriel, D. Charissoux, V. Voloshinov, and V. Molchanov, J. Lightwave Technol. **20**, 892 (2002).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1480–1485. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 123–129.

EXPERIMENTAL INSTRUMENTS AND TECHNIQUES

Elastically Deformed Thin-Walled Solar Energy Concentrators

Yu. P. Rylov

Research Institute of Electromechanics, Istra, Moscow oblast, 143500 Russia

e-mail: rprom@istra.net.ru

Received November 5, 2002; in final form, May 16, 2003

Abstract—The cheapening of solar energy reception and concentration is of primary importance for its effective use. Conventional preformed parabolic reflectors are too bulky and heavy for the shape of their surface to be maintained by load-bearing frameworks. The fabrication of elastically deformed thin-film or thin-sheet reflectors is considered in terms of the classical elasticity theory. It is shown that the solar energy concentration ratio achievable in receivers made by this technology is comparable to that in present-day collectors used in solar power plants; however, collectors made by the new technology may be much cheaper. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The startup of nine high-power (from 30 to 80 MW) solar electric generating systems (SEGSs) in the last decade in California has stimulated interest in solar power. The cost of SEGS-generated energy has approached that of energy from thermal stations. However, high capital costs, up to 50% of which are spent on solar energy concentrators and receivers, still do not allow SEGSs to compete with nuclear and thermal stations on nonrenewable fuels, although the latter are a certain environmental hazard. Below, we invoke the theory of elasticity [1] to analyze elastically deformed thin-film (thin-sheet) solar concentrators, which are cheaper and less material-consuming.

STATEMENT OF THE PROBLEM

The surfaces of solar concentrators are currently composed of rigid parabolic metallic or glass fragments [2–4]. At the same time, near-parabolic or near-cylindrical surfaces may also be produced by deforming thin films or sheets. This work is aimed at estimating (i) the deviation of the reflecting surface of elastically deformed sheets or metallic films from the ideal parabolic or cylindrical shape and (ii) a related change in the solar energy concentration ratio in the reflector–receiver system.

ELASTICALLY DEFORMED PARABOLIC CONCENTRATORS

Let us consider three ways of thin-sheet deformation that give a shape approaching the shape of the standard parabolic concentrators most closely: (i) a sheet is fixed at one end and loaded by uniformly distributed transverse moments with a linear density M_z at the other end, (ii) a sheet is fixed at one end and loaded by uniformly distributed transverse forces with a density K_y, and (iii) a sheet is hinged at both ends and longitudinally loaded by forces with a density K_x that deform the plane shape of the sheet. Let the sheet be fixed along the z axis, the neutral line of the undistorted sheet be aligned with the x axis, and bending occur along the y axis. Then, according to [1], the respective differential equations for the line of bend for the three variants of deformation will have the form

$$\frac{D}{a}\frac{d^2y}{dx^2} - M_z = 0, \tag{1}$$

$$\frac{D}{a}\frac{d^3y}{dx^3} - K_y = 0, \tag{2}$$

$$\frac{D}{a}\frac{d^4y}{dx^4} - K_x\frac{d^2y}{dx^2} = 0, \tag{3}$$

where $D = Eh^3a/12(1-\mu^2)$ is the cylindrical stiffness of the sheet, h is the sheet thickness along the y axis, a is the size of the sheet in the z direction, μ is the Poisson's ratio, and E is the Young's modulus of the sheet.

The solution of Eq. (1) yields

$$y = \frac{M_z a x^2}{2D} = \frac{x^2}{2p_1}. \tag{4}$$

This is a canonical equation of parabola $y_p = x_p^2/2p$ in the coordinates (x_p, y_p) with the parabolic vertex at the origin ($p = p_1 = D/M_z a$ is the focal parameter). Solving Eq. (2) yields

$$y = \frac{x^2}{2p_2}\left(1 - \frac{1}{3}\frac{x}{l}\right), \tag{5}$$

where $p_2 = D/K_y a l$ and l is the sheet length along the x axis.

1063-7842/03/4811-1480$24.00 © 2003 MAIK "Nauka/Interperiodica"

At small x, line of bend (5) closely approximates a parabola. To retain the parabolic shape, one may decrease the sheet thickness $h = f(x)$ as x grows. Solving Eq. (3), we find that the line of bend after the initial shape of the sheet has lost stability at $K_x \geq \pi^2 D/al^2$ is described by

$$y = A\sin kx, \tag{6}$$

with $k = \sqrt{K_x a/D}$ $(kl = \pi)$.

If the origin of the coordinate system (x, y) is transferred from the hinge to the center of the sheet, Eq. (6) in the new coordinate system (x_1, y_1), where $kx_1 = \pi/2 - kx$ and $y_1 = A - y$, takes the form

$$y_1 = A(1 - \cos kx_1) \approx A\left[\frac{(kx_1)^2}{2} - \frac{(kx_1)^4}{4} + \ldots\right]. \tag{7}$$

From (7), it follows that the line of bend closely approximates a parabola when kx_1 is small.

Equations (1)–(3) describe a small bend of the sheet; i.e., they are applicable when the bend y is much smaller than the sheet length: $y/l \ll 1$. In parabolic trough concentrators, the ratio of the focal length $p/2 \approx y$, which describes the "depth" of a parabola, to the length l of the sheet usually equals or exceeds 0.25. Therefore, a fill-scale concentrator may be composed of fragments that are weakly bent sheets having the same radii of curvature as the parabolic segments being implanted. As is known, the radius of curvature R_i of an ith parabolic segment related to the focal parameter p of this parabola, at a point with the coordinate y_{pi} also related to the focal parameter p, $y_{0i} = y_{pi}/p$, is given by

$$R_{0i} = \frac{R_i}{p} = (1 + 2y_{0i})^{\frac{3}{2}} = \sin^{-3} u_{0i}, \tag{8}$$

where u_{0i} is the angle between the tangent to the parabola at the point y_{0i} and the parabola axis (the y_p axis).

If the origin is coincident with the point (x_{pi}, y_{pi}) and the x axis is directed along the tangent to the parabola at this point, the tangent angle is given by

$$\begin{aligned}\Delta u_i &= u_{0i} - u_{ei} \\ &= \arcsin(1 + 2y_{0i})^{-1/2} - \arcsin(1 + 2y_{ei})^{-1/2} \\ &\approx 1/2(y_{ei} - y_{0i}) + (y_{ei}^2 - y_{0i}^2) + \ldots,\end{aligned} \tag{9}$$

where u_{0i} is defined above and u_{ei} is the angle between the tangent at the end of the segment and the y_p axis. Accordingly, the tangent angles for the first, φ_1 (Eq. (4)), and second, φ_2 (Eq. (5)), variants of loading, as well as the associated radii of curvature R_1 and R_2, are expressed as

$$\varphi_1 = \frac{dy}{dx} = \frac{x}{p_1}, \quad \frac{p_1}{R_1} = \frac{d^2y}{dx^2} = 1, \tag{10}$$

$$\varphi_2 = \frac{dy}{dx} = \frac{x}{p_2}\left(1 - \frac{x}{2l}\right), \quad \frac{p_2}{R_2} = \frac{d^2y}{dx^2} = 1 - \frac{x}{l}. \tag{11}$$

From Eqs. (8)–(11), it follows that, if the radii of curvature and the focal parameters of the parabola and sheets equal each other at the point (x_{pi}, y_{pi}) $(R_i = R_1 = R_2, p = p_1 = p_2)$, the radius of curvature R_1 of the sheet loaded by the moments M_z (see (10)) will be smaller than the radius of curvature R_i of the parabola (see (8)) at the end of this segment. At the same time, the radius of curvature R_2 of the sheet loaded by the forces K_y (see (11)) will be greater than R_i, with the end portion of the sheet being straight (unbent). Accordingly, the angle of inclination φ_1 to the abscissa axis at the end of the moment-loaded sheet will be greater than the angle Δu_i of the parabola, while the angle φ_2 for the force-loaded sheet will be smaller than Δu_i. The lines of bend for the sheets implanted by the first (curve *1*) and second (curve *2*) variants of loading, as well as the ith segment of the parabola (curve *3*), are schematically shown in Fig. 1.

According to the position of the sheet implanted and the parabolic segment, one may try various combinations of loading in order to bring the sheet shape as close to the truly parabolic one as possible. For example, near the vertex of a parabola with the focal parameter p, sheets of length l should be loaded by moments M_z in view of (4), (8), and (10):

$$\varphi_1 = \frac{l}{p_1} = \frac{M_z a\sqrt{\Delta x_{0i}^2 + \Delta y_{0i}^2}}{D}, \tag{12}$$

where $\Delta x_{0i} = x_{0i+1} - x_{0i}$, $\Delta y_{0i} = y_{0i+1} - y_{0i}$, and $p_1 = R_{i+1} = p(1 + 2y_{0i+1})^{3/2}$.

The shape of fragments that are farther from the vertex will approach the line of bend for the sheet loaded by forces K_y, according to (5), (8), and (11). For these fragments, an equation similar to (12) will be valid. The

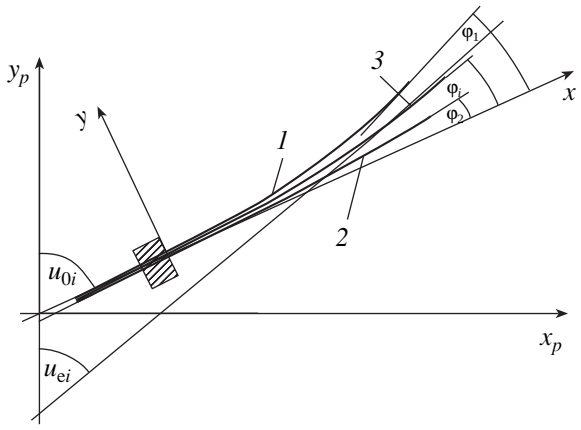


Fig. 1. Bend of plane sheets fixed at one end and loaded by (*1*) distributed moments of forces and (2) transverse forces at the other end versus (*3*) parabolic segment.

shorter the sheet, the smaller the deviation from the parabolic shape.

FILM EVACUATED CYLINDRICAL CONCENTRATORS

In the previous section, the sheet was deformed by pure bending, and tensile or compressive deformations were absent: one or both ends of the sheet either were free or freely rested on a support. The potential energy of a deformed element of the sheet was the energy absorbed in bending, and the energy of tension could be neglected. Now we will consider the other extreme case, where the potential energy of bending is small compared with the potential energy of tension. This is the case of film evacuated ferroconcrete concentrators.

Let a rectangular film be fixed between two rectilinear parallel supports and let its surfaces experience a gas pressure difference Δp. We assume that the z and x axes of the Cartesian coordinate system (x, y, z) with the origin middled between the rectilinear supports are directed, respectively, parallel and normally to them. Let the load in the z-axis direction be absent. Then, in the direction normal to the z axis, balance equations for a film element of length dl and width a will be similar to those for torsion fiber [1]:

$$\frac{d\mathbf{F}}{dl} = -\Delta\mathbf{p}a, \tag{13}$$

$$\frac{d\mathbf{M}}{dl} = [\mathbf{F}\times\mathbf{t}], \tag{14}$$

where $d\mathbf{F}$ and $d\mathbf{M}$ are the increments of the force and moment of forces due to internal stresses in a section ah at a distance dl and $\mathbf{t}$ is the unit vector in the direction to dl.

It should be noted that the statement of the problem is free of prior conditions, e.g., that the film between the supports is under tension or, conversely, has overlaps uniformly distributed along the x axis.

Let the moment of forces throughout the film and on the supports be absent: $\mathbf{M} = 0$. Then, the vector $\mathbf{t}$ is parallel to the vector $\mathbf{F}$ (see (14)) and they lie in the plane (xy). Taking into account that the gas pressure vector $\Delta\mathbf{p}$ is always perpendicular to the film surface and, hence, the vector $\mathbf{t}$ and that the tension $\mathbf{F}$ is constant throughout the film length, we find from (13) relationships for the projections of the vectors:

$$\frac{d}{dl}\left(\mathbf{F}\frac{dx}{dl}\right) = \Delta\mathbf{p}a\frac{dy}{dl}, \quad \frac{d}{dl}\left(\mathbf{F}\frac{dy}{dl}\right) = \Delta\mathbf{p}a\frac{dx}{dl};$$

hence, it follows that

$$F\frac{dx}{dl} = \Delta pay + C_1, \quad F\frac{dy}{dl} = \Delta pax + C_2. \tag{15}$$

Squaring expressions (15) and summing the results, we obtain

$$\frac{F^2}{\Delta p^2 a^2} = (x+d)^2 + (y+b)^2 = R^2, \tag{16}$$

where $d = C_2/\Delta pa$ and $b = C_1/\Delta pa$.

The constants of integration C_1 and C_2 are found as follows. Since the film is uniformly loaded by the gas pressure and the supports are arranged symmetrically about the plane (yz), the deformation of the film may be assumed to be symmetric about this plane. Then, the derivative dy/dl or the slope of the deformation line changes sign on the z axis at $x = 0$ and $y = 0$; that is, $dy/dl = 0$ and $dx/dl = 1$. Therefore, $C_1 = F$ and $C_2 = 0$ in expressions (15).

Let the film have overlaps; i.e., let its initial length l be greater than the spacing L between the supports. Then, designating the central angle in the plane (xy) as φ_0 and assuming that this angle is measured from the y axis toward the radius vector R drawn from the center of the arc to the support, we can find a relationship between the parameters that follows from the Hooke law and the geometry of deformation lines (see (16)):

$$F = Ra\Delta p, \quad \sigma = \frac{F}{ha}, \quad R = \frac{l}{2\varphi_0}\left(1+\frac{\sigma}{E}\right),$$
$$\frac{\sin\varphi_0}{\varphi_0} = \frac{L}{l\left(1+\frac{\sigma}{E}\right)} \approx 1 - \frac{\varphi_0^2}{3!} + \frac{\varphi_0^4}{5!} - \ldots, \tag{17}$$

where h is the thickness of the film.

In this special case of deformation, the film is unstressed in the z direction. If an initial stress σ_{0z} in the z direction is present and also if the gas pressure Δp generates an additional stress σ_z, the term σ/E in (17) should be replaced by the term $(\sigma - \mu(\sigma_{0z} + \sigma_z))/E$, which takes into account tensile strains in two axes. In this case, the shape of the film is defined by the superposition of two deformations: in the plane (xy) (see (17)) and in the plane (yz). In the latter plane, the parameters are related by expressions similar to (17):

$$F_z = R_2 l\Delta p, \quad \sigma_2 = \frac{F_z}{hl}, \quad R_z = \frac{a}{2\beta_0}\left(1+\frac{\sigma_z - \mu\sigma}{E}\right),$$
$$\frac{\beta_0}{\sin\beta_0} = 1 + \frac{\sigma_z - \mu\sigma}{E}, \tag{18}$$

where β_0 is the angle between the y axis and radius vector R_z in the plane (yz) and F_z is the z component of the tension due to the gas pressure Δp.

The first equations in (17) and (18) are consistent with the Laplace law, which defines capillary pressure as the superposition of pressures along the principal radii of curvature in orthogonal planes.

The reflecting surface of evacuated parabolic or cylindrical trough concentrators may be made of a metallic film, foil, metallic ribbon, or their combinations. In parabolic concentrators, the reflecting surface is composed of a parabolic lattice framework. The radius of curvature of film (ribbon) segments between straight rigid legs placed along the trough axis is kept constant and given by (16). The leg spacing depends on an allowable angular deviation of the segment at the legs. For an *i*th parabolic segment, this deviation equals the angle between the tangents to the parabola and to the arc at the $(i + 1)$th leg:

$$\delta_{i+1} = \arcsin 0.5(\Delta x_{0i}^2 + \Delta y_{0i}^2)^{\frac{1}{2}} - \arcsin(1 + 2y_{0i+1})^{-\frac{1}{2}} + \arctan\frac{\Delta x_{0i}}{\Delta y_{0i}}. \quad (19)$$

Here, it is assumed that $M = 0$ at the legs. For a metallic ribbon, $M > 0$ at intermediate legs; therefore, the deviation calculated by (19) will be somewhat overestimated.

Under service conditions, the pressure difference Δp may be very small ($\Delta p \approx 1.5$ kPa), which provides the stability of the film shape at a most plausible flow velocity $\nu = 8$ m/s, which generates an extra pressure $\Delta p_1 = 40$ Pa. The reflecting surface must be covered by a protective metal oxide (like Al_2O_3) or polymer (mylar or kapton) coating, which also may act as a sealer in structures with metal ribbons or sheets.

CONCENTRATOR DESIGN AND TECHNOLOGY

Structurally, an elastically deformed parabolic trough concentrator is a metal frame on which sheets or ribbons covered by a reflecting coating are mounted. The frame is assembled of arched ribs to which straight stringers are fastened. The coordinates of the stringers coincide with those of some points located on a given parabola with a focal parameter p. The film-bearing surfaces of the stringers and the tangents to the parabola at these points lie in the same planes. One end of the sheet or ribbon is rigidly fixed on the bearing surface of each of the stringers, while the other end is mounted with a slide fit on the neighboring stringer. The stringer spacing is equal to the width of the ribbon: $l = (0.1–0.2)p$. A correcting stringer may be placed between bearing stringers at a distance $x_0 = (0.6–0.8)l$ from the fixed end, where the deviation Δ of the deformed sheet from the parabolic shape is the greatest: $\Delta = (0.5–1.0) \times 10^{-3}p$. The angular inaccuracy in the area of greatest deviation is no higher than several angular minutes, and the angular inaccuracy averaged over the entire surface between stringers is $\delta \approx 2'$. The analysis of the elastic line of the sheet and its deviation from the parabolic shape is described below.

An evacuated cylindrical trough film concentrator is a hermetically sealed arched cavity. On the inside, it is bounded by the reflecting surface of the film rested on straight stringers; on the outside, by the U-shaped rigid metallic surface mounted on the stringers and ribs. The shape of the reflecting surface of the metallic film depends on the gas pressure and the arrangement of supports along the circumference. The pseudofocal length of the cylindrical reflector is kept constant by controlling the gas pressure inside and outside the cavity with a focus position sensor [5].

Let us perform comparative analysis of these two designs. The parabolic concentrator used in the SEGS-VIII has the focal parameter $p = 3m$ and is made of 3-mm-thick self-supported square glass mirrors of side $a = 0.5p$. The width and length of the concentrator accommodate four and six to seven mirrors, respectively. An equivalent concentrator can be made of elastically deformed metallic (aluminum alloy or steel) ribbons 0.1 mm (or more) thick, $(0.1–0.2)p$ wide, and as long as the concentrator module. With such a geometry, the number of reflecting elements is the same as in the previous example by order of magnitude. However, the weight of the reflecting surface of the thin-walled concentrator is one order of magnitude lower. Because of this, the frame of the bearing pivoted system may be made much lighter than that of the system with self-supported mirrors. Under rotation, gas-dynamic loads introduce nearly the same optical errors as the weight of 3-mm-thick corner-anchored glass mirrors [2].

CALCULATION

The shape of the elastic line of a sheet implanted into an *i*th segment of the parabolic surface is found by jointly solving two equations. The former, which is the sum of Eqs. (4) and (5),

$$y_i = \frac{x_i^2}{2p_1} + \frac{x_i^2}{2p_2}\left(1 - \frac{1}{3}\frac{x_i}{l_i}\right),$$

determines the coordinates of the elastic line under the simultaneous action of the moment M_z and force K_y. The latter allows one to find the angles of rotation of sheet sections and is the sum of Eqs. (10) and (11):

$$\varphi_i = \frac{x_i}{p_1} + \frac{x_i}{p_2}\left(1 - \frac{x_i}{2l_i}\right). \quad (20)$$

Designating the final displacement (at $x_i = l_i$) of points on the sheet as $\Delta y_i = y_i/l_i$ and the final angle of rotation of a section as $\varphi_i = \Delta u_i$, we eliminate p_1 and p_2 to obtain

$$y_i = \frac{a_i}{l_i}x_i^2 - \frac{b_i}{l_i^2}x_i^3, \qquad \varphi_i = \frac{2a_i}{l_i}x_i^2 - \frac{3b_i}{l_i^2}x_i^2, \quad (21)$$

where

$$a_i = 3\Delta y_i - \Delta u_i, \quad b_i = 2\Delta y_i - \Delta u_i.$$

Here,

$$p_1 = l_i^2/(4\Delta u_i l_i - 6\Delta y_i),$$
$$p_2 = l_i^2/(12\Delta y_i - 6\Delta u_i l_i). \tag{22}$$

Equating the final displacement of points to the difference in the ordinates and the final angle of rotation of the section, to the angle between the tangents drawn at the beginning and end of the *i*th parabolic segment, we come to equations for coefficients (21):

$$\Delta y_i = -(x_{pei} - x_{p0i})\cos u_{0i} + \frac{1}{2}(x_{pei}^2 - x_{p0i}^2)\sin u_{0i},$$

$$\Delta u_i = u_{0i} - u_{ei},$$

where x_{p0i} and x_{pei} are the abscissas of the beginning and end of the *i*th parabolic segment to be replaced in the coordinate system (x_p, y_p) of a parabola with the vertex coinciding with the origin and u_{0i} and u_{ei} are the angles the tangents at the beginning and end of the *i*th segment make with the coordinate axis.

Moments and forces that must be applied to the end of the sheet in accordance with Eqs. (4) and (5) and the parameters p_1 and p_2 in (22) are given by

$$M_z = \frac{D}{p_1 a} = \frac{2D}{l_i^2 a}(2\Delta u_i l_i - 3\Delta y_i),$$

$$K_y = \frac{D}{p_2 a l_i} = \frac{6D}{l_i^3 a}(2\Delta y_i - \Delta u_i l_i).$$

The deviation of the line of bend from the *i*th parabolic segment can be found by expressing the coordinates of the *i*th segment in the coordinate system of the sheet and determining the difference in the coordinates of these curves and associated tangent angles in the two coordinate systems. In the coordinate system of the sheet, the coordinates x_p, y_p of the parabola have the form

$$x_{pi} = (x_p - x_{p0i})\sin u_{0i} + (x_p^2 - x_{p0i}^2)\frac{\cos u_{0i}}{2p},$$

$$y_{pi} = -(x_p - x_{p0i})\cos u_{0i} + (x_p^2 - x_{p0i}^2)\frac{\sin u_{0i}}{2p}.$$

Hence, the deviation of the line of bend of the sheet from the parabolic segment and the angular deviations of the tangents to them are

$$\delta x = x_{pi} - x_i, \quad \delta y = y_{pi} - y_i,$$
$$\delta\varphi = u_i - \varphi_i.$$

Analytically, a relationship between the coordinates is hard to establish; therefore, this problem is more easily and vividly solved graphically.

SOLAR ENERGY CONCENTRATION RATIO

Solar energy concentration in the receiver is a most important property of a solar concentrator. It is estimated by the mean concentration ratio *C*, which equals the ratio of the total sunshine power incident on the concentrator surface to the power received by the focal area of the receiver. In other terms, it is equal to the ratio between the exposed surface area S_0 of a concentrator and the focal surface area S_1 of the receiver times the reflection coefficient η_r: $C = \eta_r S_0/S_1$.

Taking into account that real surfaces of reflectors locally depart from perfect geometrical (parabolic or cylindrical) surfaces by an angle δ, one should add 2δ to the maximal angle of divergence (nonparallelism) of the solar radiation reflected, i.e., to the solid semi-angle of the Sun α_0, as follows from the laws of geometrical optics. Then, at the normal incidence of solar radiation on the exposed surface of a parabolic trough concentrator with an aperture Θ_0 and focal parameter *p*, the mean concentration ratio is given by

$$C_n = \eta_r\frac{S_0}{S_1} = \eta_r\frac{\sin\Theta_0\cos\Theta_0}{\frac{\alpha_0}{2} + 2\delta} = \eta_r\frac{\sin 2\Theta_0}{\alpha_0 + 4\delta} \tag{23}$$

for a plane receiver (like a photoelectric transducer) or

$$C_r = \eta_r\frac{S_0}{S_1} = \eta_r\frac{\sin\Theta_0}{\left(\frac{\alpha_0}{2} + 2\delta\right)\left(\frac{\pi}{2} + \Theta_0\right)} = \eta_r\frac{4\sin\Theta_0}{(\alpha_0 + 4\delta)(\pi + 2\Theta_0)} \tag{24}$$

for a cylindrical receiver (like a heat exchanger tube).

Figure 2 shows dependences (23) and (24) for $a_0 = 32'$ and various δ.

Based on the conclusions drawn in [6, 7] for spherical concentrators, one may show that the mean concentration ratio in the focal plane *R*/2 of a cylindrical trough concentrator, where *R* is the radius of the cylinder, is given by

$$C_{nc} = \eta_r\frac{\sin\varphi_m}{\tan\left(2\varphi_m + \frac{\alpha_0}{2}\right)\left(\cos\varphi_m - \frac{1}{2}\right) - \sin\varphi_m}, \tag{25}$$

where $\varphi_m \approx \Theta_0/2 + \Theta_0^3/16$ is the central angle of the cylinder.

Function (25) is shown in Fig. 2 (curve *6*). It is seen that $C_{nc} \ll C_n$. The concentration ratio of a cylindrical reflector can be raised if the plane of reception is shifted

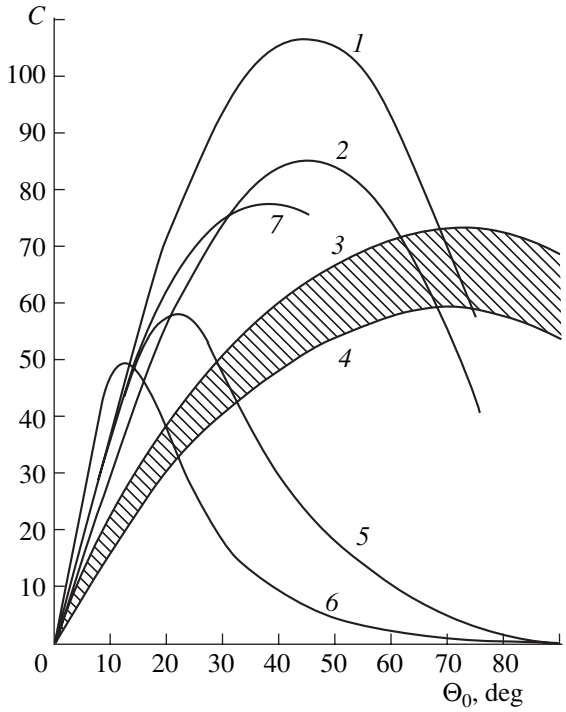


Fig. 2. Mean concentration ratios for (*1–4*) thin-walled parabolic trough concentrators and (*5–7*) film cylindrical trough concentrators. (*1*) $\delta = 0$, plane receiver; (*2*) $\delta = 2'$, plane receiver; (*3*) $\delta = 0$, cylindrical receiver; (*4*) $\delta = 2'$, cylindrical receiver; (*5*) the plane of reception at a distance of $R/2 - \Delta l$; (*6*) the focal plane at a distance of $R/2$; and (*7*) receiver with contrareflectors.

from the plane $R/2$ toward the reflector by a distance [6–8]

$$\Delta l = 0.45\frac{R}{2}\left(\frac{\Theta_0}{2} + \frac{1}{16}\Theta_0^3\right)^2.$$

In this case, the concentration ratio increases to

$$C_{nw} = \eta_r \frac{\sin 2\Theta_0}{\alpha_0 + \frac{1}{16}\Theta_0^3}. \qquad (26)$$

Dependence (26) is shown in Fig. 2 (curve *5*). The maximal value of C_{nw} is higher than C_{ne}, $C_{nw} \approx 1$, $2C_{ne}$, but smaller than the concentration ratio for parabolic reflectors (Fig. 2; curves *1*, *2*). The concentration ratio of parabolic reflectors can be reached if V-shaped contrareflectors [5] with a reflection coefficient η_r' are installed on the receiver. Then, the concentration ratio is expressed as

$$C_k = \eta_r \frac{4[\sin\varphi_k + \eta_r'(\sin\varphi_m - \sin\varphi_k)]}{\alpha_0}, \qquad (27)$$

where $\varphi_k \le 6°$ is the angle that makes a parabolic arc virtually indistinguishable from a circular one. Dependence (27) is shown in Fig. 2 (curve *7*) for $\eta_r = \eta_r' = 0.9$.

The best parabolic trough concentrations (LUZ Co.) are assembled from 3-mm-thick self-supported rigid mirror-polished sheets with a reflection coefficient $\eta_r = 0.94$ and offer a concentration ratio $C = 0.61$–0.71 [9]. These values are comparable with the ultimate value for cylindrical receivers (Fig. 2; curves *3*, *4*). The concentration ratio of elastically deformed thin-sheet concentrators (Fig. 2, curve *7*) also approaches this level.

At the same time, the cost and weight of concentrators made of thin sheets or films (0.1 to 0.3 mm thick) are two to five times lower than those of self-supported rigid reflectors made of sheets thicker than 3 mm [9].

CONCLUSIONS

(1) Solar energy trough concentrators made of elastically deformed sheets or films may offer concentration ratios approaching those of preformed rigid mirrors ($C_n \le 80\%$).

(2) Solar energy collectors consisting of elastically deformed reflecting sheets and/or films and receivers with contrareflectors are two to five times lighter and cheaper than collectors based on rigid self-supported concentrators.

REFERENCES

1. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 7: *Theory of Elasticity* (Nauka, Moscow, 1965; Pergamon, New York, 1986).
2. *Sunlight Concentrators: Collection of Articles*, Ed. by V. A. Grilikhes (Énergiya, Leningrad, 1971).
3. *Sunlight Concentrators for Photoelectrical Power Plants: Collection of Articles*, Ed. by V. A. Grilikhes (Énergoatomizdat, Leningrad, 1986).
4. D. S. Strebkov and É. V. Tver'yanovich, Teploénergetika, No. 2 (1999).
5. Yu. P. Rylov, RF Patent No. 2001117276/06(018593) (2001).
6. V. B. Veĭnberg, *Optics in Facilities Utilizing Solar Energy* (Oborongiz, Moscow, 1959).
7. O. I. Kudrin, *High-Temperature Solar Space Propulsive Plants* (Mashinostroenie, Moscow, 1987).
8. N. V. Khartchenko, *Thermische Solaranlagen* (Springer-Verlag, Berlin, 1994).
9. G. C. Goy, M. Horn, J. Lang, *et al.*, *Kostenaspekte erneuerbarer Energiequellen* (Springer Verlag, Oldenbourg, 1991).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1486–1488. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 130–132.

SHORT COMMUNICATIONS

Nonlinear Dynamics of the Evolution of Thermomagnetic Perturbations in Composite Superconductors

N. A. Taylanov and A. Élmurodov

Research Institute of Applied Physics, Tashkent State University, Universitetskaya ul. 95, Vuzgorodok, Tashkent, 700174 Uzbekistan

e-mail: taylanov@iaph.tkt.uz

Received February 21, 2003

Abstract—Nonlinear dynamics of the thermal and electromagnetic perturbations in composite superconductors is studied. © *2003 MAIK "Nauka/Interperiodica".*

It is known that, at the linear stage, the evolution of thermomagnetic perturbations exhibits exponential (increasing or decreasing) behavior [1]. This stage implies that the ohmic power is either temperature-independent or is a linear function of temperature. However, at the stage of nonlinear instability, some electrical and thermal characteristics, including the heat-release power, depend heavily on temperature. Therefore, the analysis of the critical state stability should be performed with regard for nonlinear temperature dependence of heat release.

This study is concerned with the dynamics of the thermomagnetic perturbations that develop in composite superconductors.

Let us consider the main equations that describe the evolution of thermal and electromagnetic perturbations in a superconductor in the critical state. The propagation of heat is given by nonlinear heat-conduction equation

$$\nu(T)\frac{dT}{dt} = \nabla \cdot [k(T)\nabla T] + \mathbf{j} \cdot \mathbf{E}, \qquad (1)$$

where ν and k are heat capacity and thermal conductivity, respectively.

The last term in Eq. (1) corresponds to the power of Joule heat release in the region of the normal phase. Both quantities ν and k may be considered temperature-independent if the cooler operates at a temperature close to the critical value T_c.

The evolution of electromagnetic field perturbations is described by the system of Maxwell equations

$$\mathrm{curl}\mathbf{E} = -\frac{1}{c}\frac{d\mathbf{H}}{dt}, \qquad (2)$$

$$\mathrm{curl}\mathbf{H} = \frac{4\pi}{c}\mathbf{j}. \qquad (3)$$

In order to close the system of Eqs. (1)–(3), one should establish the relation between **j**, **E**, **H**, and T. Let us assume that the values of **E**, **H**, and T lie on the so-called resistive surface

$$\mathbf{j} = \mathbf{j}(T, \mathbf{H}, \mathbf{E}). \qquad (4)$$

In a wide range of the electric-field strengths E, the current–voltage (I–V) characteristic of superconductors can be represented in the form

$$\mathbf{j}(T, E, H) = [\mathbf{j}_c(T, \mathbf{H}) + \mathbf{j}_n(E)]\frac{\mathbf{E}}{E}. \qquad (5)$$

As is well known [1], at $j \le j_c$, the vortices start their motion, which is a viscous flow across the applied transport current. Such flow is possible only for the electric fields $E \le E_f$. Here, E_f is the boundary of the linear section in I–V characteristic. In the specified region, the dependence $j_r(E)$ is linear: $j_r \sim \sigma_f E$ [2]. With variation of electric field E, the I–V characteristic rapidly becomes linear (at $E \le E_f$), while the actual values of E in superconductor meet the condition $\sigma_f E \ll j_c$; i.e., the density of the vortex current $\sigma_f E$ is small as compared to that of the critical current j_c. This circumstance enabled Bean to formulate the critical-state concept [3]. According to this concept, any effect that gives rise to electric field E results in the establishment of a critical state with density j_c. The critical state concept provides a basis for the macroscopic electrodynamics of high-current superconductors.

For superconductors, the linear dependence $j_r(E)$ holds true only in the region $E \le E_f$; outside this region, the differential conductivity σ is heavily E-dependent: $\sigma(E) = dj/dE$ (flux creep) [4]. The I–V characteristics of superconductors in the creep mode have universal character and can be described by the same formula in a wide range of temperatures and magnetic fields:

$$j(T, E, H) = j_c(T, H) + j_1(E)\ln\frac{E}{E_0}. \qquad (6)$$

Here, E_0 and j_1 are constants. The last formula implies that the differential conductivity grows with a decrease

1063-7842/03/4811-1486$24.00

in E, i.e., $\sigma(E) = j_1/E$. The causes and mechanisms behind the nonlinearity of I–V characteristics of superconductors in weak fields are discussed in detail in [2]. One possible mechanism is the thermally activated flux creep. In the range of very low fields ($E \longrightarrow 0$), the I–V characteristic may again be linear: $j_n \sim \sigma_d E$, with $\sigma_d \ll \sigma_j$.

The dependence of the critical current density on the temperature of a superconducting material in the critical state is defined by the following equation:

$$j_c(T) = j_c(T_0)\left[1 - \frac{T - T_0}{T_c - T_0}\right], \quad (7)$$

where $j_c(T_0)$ is the critical current density at T_0.

Let us consider a planar semi-infinite sample ($x > 0$) subjected to magnetic field $\mathbf{H} = (0, 0, H_e)$ that increases at a constant rate $d\mathbf{H}/dt$ = const. The vortex electric field $\mathbf{E} = (0, E_e, 0)$ arising from the magnetic flux motion is parallel to the plane of current $\mathbf{E} \parallel \mathbf{j}$ [5].

Thermal and electromagnetic boundary conditions to Eqs. (1)–(4) have the form

$$\frac{dT(0)}{dx} = 0,\ T(L) = T_0,\ \frac{dE(0)}{dx} = 0,\ E(L) = 0. \quad (8)$$

Here, L is the characteristic magnetic flux penetration depth. This quantity can easily be found from Maxwell equation (3) with the boundary conditions $H(0) = H_e$ and $H(L) = 0$ and is defined as

$$L = \frac{cH_e}{4\pi j_0}.$$

It is known [2] that the propagation of magnetic flux and heat flow is characterized by the corresponding diffusion coefficients: the magnetic diffusion coefficient

$$D_m = \frac{c^2}{4\pi\sigma_f}$$

associated with the normal currents in the resistive superconducting mode and the thermal diffusion coefficient $D_t = k/\nu$. The ratio of these quantities $\tau = D_t/D_m$ governs the dynamics of the development of thermal and electromagnetic perturbations in composite superconductors with $\tau \gg 1$. For the sake of simplicity, the following discussion is restricted to the approximation of slight heating $T - T_0 \ll T_0$. Within this approximation, the coefficients ν and k may be considered as independent of the temperature profile.

The effective conductivity of composite superconductors σ_f is appreciably higher than that of the hard superconductors. Therefore, we may assume that the induced normal current $\sigma_f E$ compensates the drop of the temperature-induced critical current $j_c(T)$ and thus hinders the magnetic flux penetration into the sample. Within such formulation of the problem, the magnetic-flux motion may be disregarded: $D_m \ll D_t$ or $\tau = D_t/D_m \gg 1$. In other words, the development of magnetic instability in composites is accompanied with "slow" perturbations with characteristic rise time $t_j \gg t_k$ or $t_j \ll t_m$, where $t_k = L^2/D_t$ and $t_m = L^2/D_t$ are the times of thermal and magnetic diffusion, respectively. Then, according to Maxwell Eqs. (2) and (3), the current density in each point of the sample remains constant $dj/dt = 0$, and the system of Eqs. (2) and (3) yields a linear relation between the temperature and the electric field:

$$T(x, t) = \frac{\sigma_f}{a}E(x, t) + T_0. \quad (9)$$

Substituting this expression into Eq. (1) yields the equation for the electric field distribution $E(x, t)$ in the form

$$\frac{dE}{dt} = D_m\frac{d^2E}{dx^2} + \frac{j_0}{E_\nu}E, \quad (10)$$

where $E_\nu = \sigma_f\nu/a$ is a constant parameter.

Performing the separation of variables in Eq. (10) and taking into account the boundary conditions, we have

$$E(x, t) = \exp\left(\left[\frac{j_0}{E_\nu} - \frac{\pi^2}{4L^2}D_t\right](t - t_0)\right)\cos\frac{\pi}{2L}x. \quad (11)$$

Now, the expression for temperature (9) becomes

$$T(x, t) = \frac{\sigma_f}{a}\exp\left(\left[\frac{j_0}{E_\nu} - \frac{\pi^2}{4L^2}D_t\right](t - t_0)\right)\cos\frac{\pi}{2L}x + T_0. \quad (12)$$

The solutions obtained above suggest that, under the condition

$$\frac{j_0}{E_\nu} - \frac{\pi^2}{4L^2}D_t \geq 0, \quad (13)$$

both the temperature and the field perturbations exponentially grow. The critical thickness value l_c corresponding to the onset of the avalanche build-up of perturbations is given by

$$l_c = \frac{\pi}{2}\left[\frac{k\sigma_f}{j_0 a}\right]^{1/2}. \quad (14)$$

The critical-field value H_j starting from which the magnetic flux penetrates into the superconductor bulk is expressed as

$$H_j = \frac{\pi^2}{2}\left[\frac{16 j_0 k\sigma_f}{ac^2}\right]^{1/2}. \quad (15)$$

It can be seen that, within the time interval

$$t_m \gg t \gg t_k, \quad (16)$$

the results obtained provide for a highly accurate description of the evolution of thermomagnetic pertur-

bations that take place in hard superconductors in the critical state.

Finally, it should be noted that the values of H_j and l_c depend heavily on the properties of the composite material. Our results are valid only if j is a linear function of E. As was mentioned above, the I–V characteristics of composite superconductors are distinctly nonlinear at small values of electric field $E < E_f$. Strong E-dependence of differential conductivity σ in this region should affect the criterion of stability of the critical state and, hence, influence the values of H_j and l_c. Such a problem goes beyond the scope of this study and presents a subject for future investigation.

REFERENCES

1. A. M. Campbell and J. E. Evetts, *Critical Currents in Superconductors* (Taylor and Francis, London, 1972; Mir, Moscow, 1975).
2. R. G. Mints and A. L. Rakhmanov, Usp. Fiz. Nauk **121**, 499 (1977) [Sov. Phys. Usp. **20**, 249 (1977)].
3. C. P. Bean, Phys. Rev. Lett. **8**, 250 (1962).
4. P. W. Anderson and Y. B. Kim, Rev. Mod. Phys. **36**, 39 (1964).
5. N. A. Taylanov, Supercond. Sci. Technol. **14**, 326 (2001).

Translated by A. Sidorova

Technical Physics, Vol. 48, No. 11, 2003, pp. 1489–1492. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 133–136.

SHORT COMMUNICATIONS

Penetration of Microwave Radiation into Water: Effect of Induced Transparency

V. Yu. Knyazev*, I. A. Kossyi*, N. I. Malykh*, and E. S. Yampolskii**

** Prokhorov Institute of General Physics, Russian Academy of Sciences, ul. Vavilova 38, Moscow, 117942 Russia*

e-mail: kossyi@fpl.gpi.ru

*** Sukhumi Physicotechnical Institute, Sukhumi, 384914 Abkhazia, Georgia*

Received December 15, 2002; in final form, March 25, 2003

Abstract—The propagation of intense microwave radiation in a water column is studied experimentally. The effect of induced transparency of water is discovered. The effect is related to the successive heating of water layers from the surface deep into the water column. The interaction of intense microwave radiation with water is simulated with account of variations in the electrophysical water parameters. The simulation results are compared with the experimental data. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The propagation of low-power microwaves through liquids obeys the Bouguer law [1]

$$I = I_0 e^{-\alpha Z}, \quad (1)$$

which is well known in linear electrodynamics. Here, α is the absorption coefficient of microwaves incident onto a half-space willed with water, I_0 is the microwave intensity, and Z is the coordinate counted from the water surface deep into water along the normal to the surface.

The coefficient α depends on neither Z nor the microwave intensity. The reciprocal of α determines the penetration depth of an electromagnetic wave into water, $d_0 \approx 1/\alpha$.

For low-power microwaves, the absorption and reflection characteristics of liquids were studied in great detail (see, e.g., [2, 3]). Here, of interest may be only the determination of the absorption and reflection coefficients for specific media for which data are either crude or lacking.

The situation is quite different when the intensity of an electromagnetic wave is so high that it changes the state of a liquid. In this case, the penetration of radiation into the liquid cannot be described by linear electrodynamics.

The objective of this paper is to investigate the penetration of high-power microwave radiation into water and its solutions.

EXPERIMENTAL SETUP AND MEASUREMENT RESULTS

A schematic of the experimental setup is shown in Fig. 1. Magnetron *1* is connected to standard waveguide *2*. The waveguide output of the magnetron is connected to two conical transitions *4* and *9* through matched flanges. Transition *9* is connected to a measuring system consisting of directional coupler *12* and calibrated

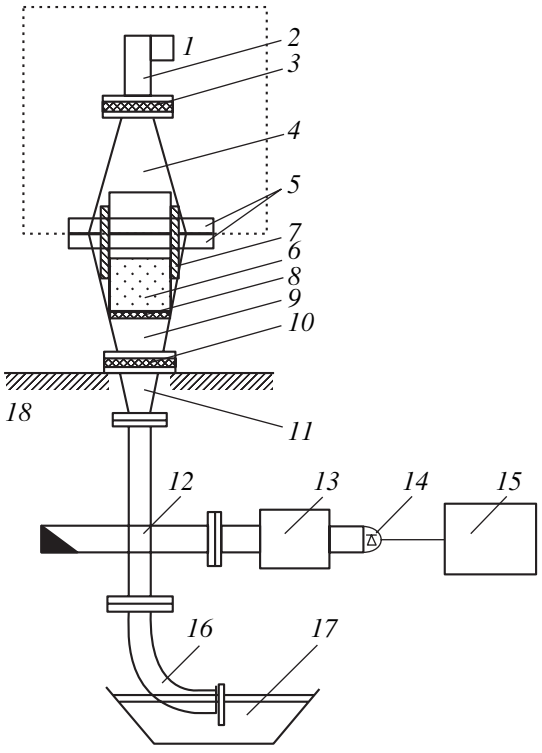


Fig. 1. Schematic of the experimental setup: (*1*) magnetron, (*2*) rectangular waveguide, (*3*, *10*) Teflon gaskets, (*4*, *9*, *11*) conical transitions, (*7*) gasket, (*12*) directional coupler, (*13*) calibrated attenuator, (*14*) detector, (*15*) oscilloscope, (*16*) waveguide flexure, (*17*) water cell (a component of the matched load), and (*18*) base of the low part of the setup.

1063-7842/03/4811-1489$24.00 © 2003 MAIK "Nauka/Interperiodica"

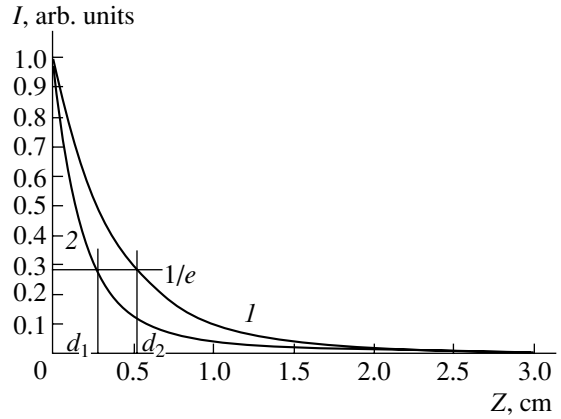


Fig. 2. Penetration of low-power microwave radiation into (*1*) water and (*2*) water solution of salt.

attenuator *13* connected to detector *14*. A signal from the detector is fed to oscilloscope *15*.

The upper part of the setup (enclosed by a dotted line) has the connection in flanges *5* and can be lifted and taken aside. This allows us to place cylindrical glass cell *6* on Teflon bottom *8*. The cell is filled with a fixed amount of water, through which microwave signals pass.

The magnetron generated a train of microwave pulses with a duration of about 8 ms and repetition rate of 50 Hz. The average microwave power was 480 W, and the peak power attained 1.8 kW. The magnetron wavelength was 12.4 cm. Copper gasket *7* was placed around a water cell in order to prevent the direct penetration of the magnetron radiation into the measuring system.

The experimental setup makes it possible to measure the absorption coefficient and investigate the penetration of microwaves into a liquid medium both under conditions corresponding to the Bouguer law [relation (1)] and when linear electrodynamics is inapplicable because of the change in the electrical properties of the irradiated liquid under the action of intense microwave radiation. One regime was changed to another by varying the irradiation time t of the liquid. At short irradiation times ($t < 1$ s), the propagation of electromagnetic waves is consistent with that described by linear electrodynamics; hence, this regime can be classified as irradiation by low-power microwaves. When the irradiation time is long enough ($t > 10$ s), the character of microwave penetration into the liquid changes radically. This regime can be classified as the propagation of high-power microwave radiation.

First, we consider the regime of irradiation by low-power microwaves. The measurements were carried out with distilled water, whose conductivity is negligibly small; tap water with a conductivity of 380 μS/cm; and water solution of salt with a conductivity of 2000 μS/cm.

Figure 2 shows the measurement results. Curve *1* corresponds to the measurements with distilled and tap water (in both cases, the results completely coincide). The absorption coefficient is $\alpha = 2$ cm^{-1}. Accordingly, the depth d_2 at which the intensity decreases by a factor of e is 5 mm. When we added salt in water and increased the solution conductivity to 2000 μS/cm, the absorption coefficient increased to 4 cm^{-1} (curve *2*) and the absorption depth d_1 was 2.5 mm.

The regime in which the propagating microwave radiation substantially changed the dielectric properties of the liquid (the regime of nonlinear electrodynamics) was studied in the same experimental device.

Magnetron *1* was switched on for a time that was substantially longer than 1 s (for 10, 20,…, 120 s). Then, the upper part of the setup was taken aside and the temperature in several water layers was measured. For this purpose, a small-size thermocouple immersed in the water column was displaced along the column axis (from the water surface to the cell bottom) and the temperature was measured in different layers (the total measurement time was several seconds; i.e., the temperature was measured over the time interval during which the axial temperature profile remained essentially unchanged).

After every run of the measurements, the water in the cell was mixed by a dielectric rod and the average temperature was measured. From this value, we calculated the power absorbed by the water column by the formula

$$I_{ab} = \rho c_p (T_1 - T_0)/t, \qquad (2)$$

where T_1 and T_0 are the final and initial temperatures, respectively; c_p is the water specific heat; and ρ is the water mass density.

For any irradiation time t, the mean absorbed power was about 480 W. The power passed through the water was measured by detector *14*.

The measurement results are shown in Fig. 3. The set of the experimental curves demonstrates how the temperature of the water column irradiated by high-power microwaves varies with time at different distances Z from the water surface. The time of the temperature propagation by heat conduction was substantially longer than the irradiation time (this was checked directly by placing an ohmic heater into water near its surface). Therefore, the time evolution of the temperature propagation along the Z axis reflects the time evolution of the microwave energy penetration into the water column. The observed effect of microwave penetration into water to depths substantially larger than the absorption depth d_2, corresponding to the regime of linear electrodynamics (according to Fig. 2, this depth is about 5 mm for low-conductivity water), can be classified as the effect of the induced transparency of water under the action of high-power radiation. The propagation velocity of the "transparency wave" (determined

from the curves in Fig. 3) as a function of Z is shown in Fig. 4. The velocity is maximal (on the order of 0.1 cm/s) near the water surface and decreases with Z.

DISCUSSION OF THE EXPERIMENTAL RESULTS

Thus, the effects of the induced transparency of water under the action of intense microwaves have been discovered. Microwave radiation absorbed in a narrow surface layer with a thickness d_0 determined by the Bouguer law [see (1)] heats this layer. The absorption coefficient of microwaves α is a function of the water temperature T. The temperature dependence of this coefficient for low-conductivity water (distilled or tap water) is given in the table (the data are taken from [2]). It can be seen that the absorption coefficient decreases from 2.2 to 0.4 cm^{-1} as the temperature T increases from 2 to 65°C. Hence, the heating of the surface layer is accompanied by the increase in its transparency, and the radiation penetrates deeper into water. As a result, the penetration depth progressively increases as successive water layers are heated one after another.

The propagation velocity v_z of the transparency wave can be estimated from the obvious approximate relation

$$v_z \cong I_0 \exp(-\alpha_T Z)/\rho c_p \Delta T, \qquad (3)$$

where I_0 is the intensity of microwave radiation incident onto a water column, α_T is the absorption coefficient in the heated (transparent) layer (according to [2], we take $\alpha_T \approx 0.4$ cm^{-1} at $T_h \cong 65$°C), and $\Delta T = T_h - T_0$ is the difference between T_h and the initial water temperature.

The propagation velocity of the transparency wave estimated by Eq. (3) agrees with the experiment results in both the order of magnitude and the character of its decrease deep into the water column (Fig. 4).

The experimental results were compared with computer simulations performed by the following algorithm. The water column was divided into identical layers with a thickness of $d = 1$ mm. The time step was chosen to be equal to $\Delta t = 1$ s. It was assumed that the axial profile of the microwave intensity did not change within this time interval. The intensity profile was recalculated at each time step. Assuming that the initial temperature was the same for each layer (consequently, the absorption coefficient α was assumed to be constant along the water column at the first time step Δt_1), we calculated the intensity profile by the formula

$$I_n = I_{n-1} \exp(\alpha d_n),$$

where I_n is the intensity at the entrance to the layer, n is the layer number, and d_n is the layer thickness. The quantity $\Delta I = I_{n-1} - I_n$ is the power absorbed in the nth layer. Using the relation $\Delta I = \rho c_p (T_1 - T_0)/\Delta t_1$, we calculated the temperature distribution over the layers after the first time step (Δt_1). This temperature distribution corresponded to a new axial intensity profile calculated using the dependence $\alpha(T)$ taken from [2]. This

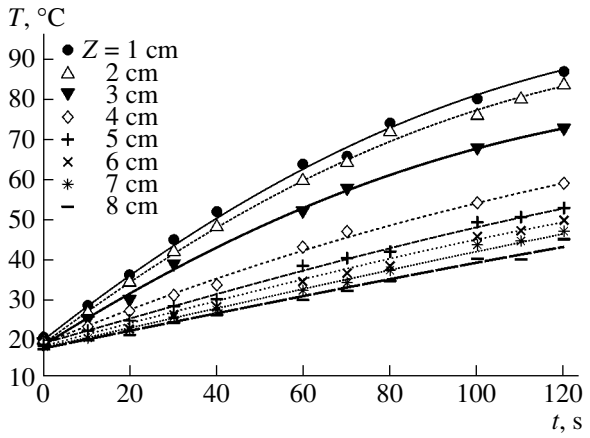


Fig. 3. Temperature of the water column at different depths vs. irradiation time.

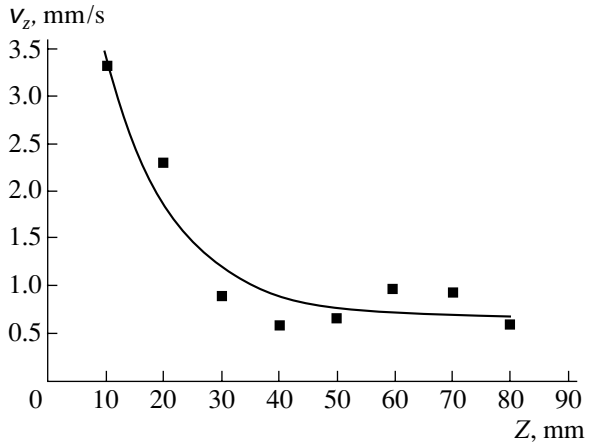


Fig 4. Propagation velocity of the transparency wave in a water column vs. depth Z.

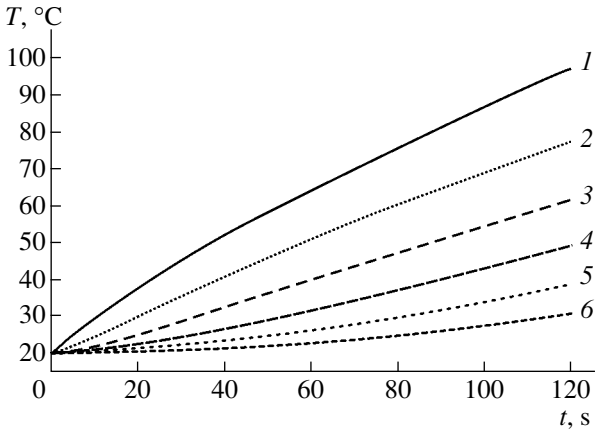


Fig. 5. Calculated temperature of the water column at different depths vs. irradiation time.

Table

T, °C	α, cm^{-1}
2	2.2
6	1.66
14	1.38
25	0.82
36	0.64
46	0.46
56	0.42
65	0.40
75	0.35
85	0.30
95	0.30

distribution was assumed to be unchanged throughout the second time interval Δt_2. Using this distribution, we calculated the temperatures of all the layers after the second time step Δt_2. From this temperature distribution, we again found a new axial intensity profile after the second time interval Δt_2. Thus, the procedure was repeated step by step many times and a series of curves was plotted as is shown in Fig. 5. One can see that the calculated curves are similar to the experimental curves shown in Fig. 3. A certain difference can be explained by the fact that the time step in simulations was not small enough; we also ignored the cooling of the upper water layers, evaporation, etc.

Experiments with water solutions of salt showed that, in comparison with the distilled or tap water, the character of penetration of both low- and high-power microwaves into the solutions begins to change starting from conductivities as high as 1500–2000 μS/cm (see, e.g., Fig. 2, curve *2*). This fact implies that the induced transparency observed in our experiments will apparently take place not only in pure water, but also in a wide class of waste waters and water solutions.

CONCLUSIONS

The effect of the induced transparency of water under the action of high-intensity ($I_p \approx 5$ W/cm^2) continuous (or quasi-continuous) microwave radiation ($\lambda \cong 12$ cm) has been discovered. It is shown that the penetration depth of high-power microwaves substantially increases as compared to low-power microwave radiation (which penetrates into water no deeper than 5 mm) because of the generation of a transparency wave, whose velocity attains 10^{-1} cm/s near the surface and decreases to $\sim 10^{-2}$ cm/s as the wave propagates along the water column.

The results obtained can be used in various technologies and industrial processes involving microwave heating of liquid flows (purification of waste water, sterilization of liquid food, etc.).

ACKNOWLEDGMENTS

We thank É.M. Barkhudarov, S.I. Gritsinin, and K.F. Sergeĭchev for their continuing interest in our study and fruitful discussions.

REFERENCES

1. I. E. Tamm, *Principles of Electricity Theory* (Nauka, Moscow, 1976).
2. G. Puschner, *Microwave Heating* (Énergiya, Moscow, 1968).
3. I. L. Khabibulin, Inzh.-Fiz. Zh. **73**, 832 (2000).

Translated by E. Satunina

Technical Physics, Vol. 48, No. 11, 2003, pp. 1493–1495. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 137–139.

SHORT COMMUNICATIONS

Thermally Stimulated Switching in Electroactive Polymer Films

A. F. Ponomarev*, V. A. Krasilnikov*, M. Vasiliev*, and A. N. Lachinov**

** Birsk State Pedagogical Institute, Birsk, 452453 Russia*

*** Institute of Molecular Physics, Ufa Scientific Center, Russian Academy of Sciences, * Ufa, 450075 Bashkortostan, Russia*

e-mail: lachinov@anrb.ru

Received December 24, 2002; in final form, April 23, 2003

Abstract—Electronic switching in poly(diphenylenephthalide) due to the thermal ionization of electron traps is studied. The method of thermally stimulated current and the method of thermally stimulated depolarization current are used. A correlation between spectra obtained by the two methods is established. The results are interpreted by using the temperature dependence of IR absorption spectra taken under conditions similar to those under which thermally stimulated phenomena are observed. Comparing data obtained, we assign electron traps to specific molecular groups of the polymer and infer the importance of the surplus space charge critical concentration for the onset of the high-conductivity state in the polymer. © *2003 MAIK "Nauka/Interperiodica".*

The phenomenon of charge instability discovered in several electroactive polymers [1–4], which causes an anomalously high local conductivity, is of great theoretical and applied interest. Yet, a generally accepted mechanism behind this phenomenon is lacking despite a variety of available models [5–8].

Available experimental data for the switching effect in metal–polymer–metal systems, which was initiated either by applying pressure [9] or electric field [10] or by varying the boundary conditions [11], suggest that the transition to the high-conductivity state follows the stage of nonequilibrium space charge accumulation. The variation of the electron energy spectrum in polymer films as a result of charge injection has been theoretically predicted and experimentally confirmed elsewhere [12, 13].

However, injection is not the only way of changing the space charge in polymer films. For example, the space charge concentration was varied by irradiation of the film by an electron beam [14].

The possibility of measuring the thermally stimulated current (TSC) has been demonstrated in [15, 16]. However, a reliable correlation between electron traps and the molecular structure of a polymer cannot be found from these data.

In this work, we study the TSC spectra and the thermally stimulated depolarization (TSD) spectra in polymer films for various film thicknesses and rates of change of temperature. Another goal is to find a relation between the thermally stimulated phenomena and electron switching effect.

The object of investigation was polymers from the class of poly(phthalidylidenearylene)s [17], in which intriguing effects related to the transition to the high-conductivity state have been observed [18, 19]. The sample, a polymer film sandwiched in metal films, was applied on a polished glass substrate with a planar electrode [11]. The TSC and TSD spectra were taken by the standard techniques [20].

Typical measurements of the thermally stimulated currents passing through the polymer film at a given rate of change of temperature are given in Fig. 1. Three portions can be distinguished in this figure. In the first one, 150–250 K, small-amplitude current fluctuations are observed. In the second portion (250–300 K), the slope of the curve $I(T)$ increases, the fluctuation amplitude grows, and a local current maximum is observed. In the third portion (>300 K), the current through the sample rises in an avalanche-like manner and the sample passes to the high-conductivity state. As the temperature grows further, the polymer conductivity remains almost constant. In the thinner polymer film (Fig. 1, curve *2*), the current increases insignificantly throughout the temperature range and the temperature of transition to the high-conductivity state slightly lowers. For a lower rate of heating (curve *3*), the transition to the high-conductivity state does not occur. Thermally stimulated processes in these polymer films may be associated not only with the ionization of electron states in the energy gap but also with a change in the mobility of molecular fragments in the range of characteristic temperatures.

To substantiate the latter reason, we took the depolarization current spectra from the same samples (Fig. 2). Here, three portions similar to those in the temperature dependence of the TSC (Fig. 1) are also seen.

From Figs. 1 and 2, it follows that, at low temperatures (below 250 K), the mobility of molecular chains,

1063-7842/03/4811-1493$24.00

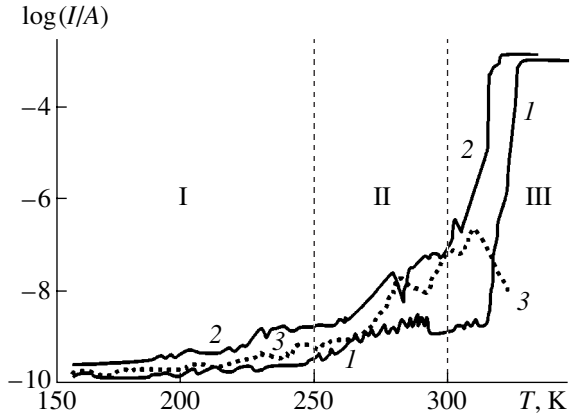


Fig. 1. Temperature dependence of the TSC. The rate of heating is 5 K/min. The film thickness if (*1*) 1.5 and (*2*) 1 μm. Curve *3* shows the behavior of sample *2* for a rate of heating of 1 K/min.

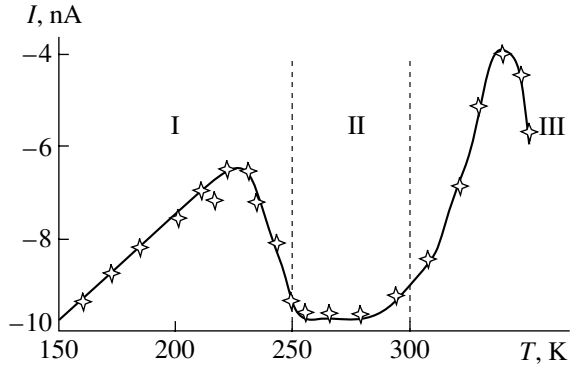


Fig. 2. Temperature dependence of the TSD current. The rate of heating is 5 K/min.

which is clearly seen in the TSD spectrum, makes a minor contribution to the TSC. At temperatures between 250 and 300 K, the TSD signal is virtually absent and the TSC spectrum exhibits a local maximum. Presumably, this indicates the ionization of electron states unrelated to mobile fragments of molecular chains. The best coincidence between the spectra is observed at temperatures above 300 K and a low rate of change of temperature. As follows from the temperature dependence of the IR absorption spectra for poly(phthalidylidenearylene)s [15], the modification of the spectra are the greatest near $\nu \sim 1600$ cm^{-1} at temperatures close to 250 and 300 K. In this range, a 1595–1608 cm^{-1} doublet is observed, which is typical of vibrating C=C bonds in the phenyl ring of a side phthalide fragment [21–23]. The observed variation of the IR absorption spectrum for poly(phthalidylidenebiphenylilene) is apparently related to the vibration of these bonds.

From the aforesaid, one may conclude that the "high-temperature" type of thermally stimulated depolarization is due to the thermal activation of a side phthalide fragment of a macromolecule. This process is accompanied by the ionization of electron traps; therefore, the TSC spectrum has a peak. When the rate of heating is high, the film passes to the high-conductivity state at the temperature of the TSC maximum.

As follows from our results, the transition to the high-conductivity state in the polymer film results in anomalously high intramolecular excitation, which shows up as an increase in the amplitude of the macromolecule vibrational spectral lines. This excitation takes place in a side phthalide fragment. It is accompanied by the occurrence of peaks in both the TSC and TSD spectra. Such a correlation can be explained in terms of quantum-mechanical calculations of the electron state density and also by taking into account the optimized configuration of a poly(diphenylenephthalide) molecule (the optimization was accomplished in [24]). It was shown in [24] that the capture of an excess electron by a side phthalide fragment may carry the molecule to a new energetically stable state of another configuration with the formation of deep electron states in the forbidden energy range. It seems that a sharp increase in the intensity of vibrational spectrum lines for a side phthalide fragment of the polymer reflects the transition of the molecule to a new state via the capture of an electron by a molecular trap. Therefore, the temperature interval where the IR spectrum lines have maximal amplitudes correlates well with temperature intervals II and III in the TSC and TSD spectra.

Thus, it was experimentally shown that electronic switching from the low- to the high-conductivity state in poly(diphenylenephthalide) films is due to the thermal ionization of deep traps. As a result, the amplitude of intramolecular vibrations in side phthalide fragments grows substantially. This means that such an electronic transition leads to the reconfiguration of the macromolecule's electronic spectrum.

REFERENCES

1. L. F. Pender and R. G. Fleming, J. Appl. Phys. **46**, 3426 (1975).
2. N. S. Enikolopyan, Yu. A. Berlin, S. I. Beshenko, and V. A. Zhorin, Pis'ma Zh. Éksp. Teor. Fiz. **33**, 508 (1980) [JETP Lett. **33**, 492 (1980)].
3. V. M. Tuchkevich and A. N. Ionov, Pis'ma Zh. Tekh. Fiz. **16** (16), 90 (1990) [Sov. Tech. Phys. Lett. **16**, 638 (1990)].
4. A. N. Lachinov, A. Yu. Zherebov, and V. M. Kornilov, Pis'ma Zh. Éksp. Teor. Fiz. **52**, 742 (1990) [JETP Lett. **52**, 103 (1990)].
5. L. N. Grigorov, Pis'ma Zh. Tekh. Fiz. **17** (10), 45 (1991) [Sov. Tech. Phys. Lett. **17**, 368 (1991)].
6. A. M. Elyashevich, A. A. Kiselev, A. V. Liapzev, *et al.*, Phys. Lett. A **156**, 111 (1991).
7. D. M. Eagles, Physica C **225**, 222 (1994).
8. O. A. Ponomarev and E. S. Shikhovtseva, Zh. Éksp. Teor. Fiz. **107**, 637 (1995) [JETP **80**, 346 (1995)].

9. A. N. Lachinov, A. Yu. Zherebov, and V. M. Kornilov, Synth. Met. **44**, 111 (1991).
10. A. N. Lachinov, A. Yu. Zherebov, and V. M. Kornilov, Zh. Éksp. Teor. Fiz. **102**, 187 (1992) [Sov. Phys. JETP **75**, 99 (1992)].
11. V. M. Kornilov and A. N. Lachinov, Zh. Éksp. Teor. Fiz. **111**, 1513 (1997) [JETP **84**, 833 (1997)].
12. C. B. Duke and T. J. Fabish, Phys. Rev. Lett. **37**, 1075 (1976).
13. T. J. Fabish, H. M. Saltsburg, and M. L. Hair, J. Appl. Phys. **47**, 930 (1976).
14. V. M. Kornilov and A. N. Lachinov, Pis'ma Zh. Éksp. Teor. Fiz. **61**, 504 (1995) [JETP Lett. **61**, 520 (1995)].
15. A. N. Lachinov, A. Yu. Zherebov, and M. G. Zolotukhin, Synth. Met. **59**, 377 (1993).
16. A. Zherebov, A. Lachinov, and V. Kornilov, Synth. Met. **84**, 917 (1997).
17. S. R. Rafikov, G. A. Tolstikov, S. N. Salazkin, *et al.*, USSR Inventor's Certificate no. 743989, *Byull. Izobret.*, 1981, no. 20.
18. A. N. Ionov, A. N. Lachinov, M. M. Rivkin, *et al.*, Solid State Commun. **82**, 609 (1992).
19. A. N. Ionov, A. N. Lachinov, and R. Rench, Pis'ma Zh. Tekh. Fiz. **28** (14), 69 (2002) [Tech. Phys. Lett. **28**, 608 (2002)].
20. V. N. Vertoprakhov and E. G. Sal'man, *Thermally Stimulated Currents in Inorganic Materials* (Nauka, Novosibirsk, 1979).
21. S. N. Salazkin, S. R. Rafikov, G. A. Tolstikov, *et al.*, Dokl. Akad. Nauk SSSR **262**, 355 (1982).
22. S. N. Salazkin and S. R. Rafikov, Izv. Akad. Nauk Kaz. SSR, Ser. Khim., No. 5, 27 (1981).
23. M. G. Zolotukhin, A. A. Panasenko, V. S. Sultanova, *et al.*, Makromol. Chem. **186**, 1747 (1985).
24. N. Johansson, A. N. Lachinov, S. Stafstrom, *et al.*, Synth. Met. **67**, 319 (1994).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, pp. 1496–1498. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 11, 2003, pp. 140–142.

SHORT COMMUNICATIONS

Effect of the Anharmonic Phase Dependence of the Supercurrent on *I–V* Hysteresis in a Josephson Junction

I. N. Askerzade

Institute of Physics, National Academy of Sciences of Azerbaijan, pr. Dzhavida 33, Baku, 370143 Azerbaijan
Department of Physics, Ankara University, 06100 Tandogan, Ankara, Turkey
e-mail: solstphs@physics.ab.az

Received January 24, 2003; in final form, April 4, 2003

Abstract—The effect of the anharmonic phase dependence of the supercurrent on *I–V* hysteresis in a Josephson junction is studied in terms of a modified resistive model. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

In [1], a Josephson junction with an antiferromagnetic spacing of thickness d between superconductors was reported. As the thickness of the antiferromagnetic spacing decreases, the current–phase dependence becomes nonsinusoidal. Calculations show [2] that the distortion of the current–phase dependence is estimated through the parameter $\delta = d/d_c$, where $d_c = \pi\xi_a$ is a critical thickness and ξ_a is the antiferromagnetic coherence length. Note that anharmonicity in the current–phase dependence arises in SNS or SINS systems (S, superconductor; N, normal metal; and I, insulator) based on low-temperature superconductors at room temperature [3, 4] and in superlattices consisting of alternating insulating and superconducting layers [5]. The order parameter symmetry in high-temperature superconductors greatly influences the value and shape of the supercurrent in Josephson junctions. It has been shown [6] that the sign of the Josephson tunnel current in d-wave superconductors depends on the order parameter orientation relative to the junction plane [7]. The properties of various Josephson junctions based on d-wave superconductors have been covered elsewhere [8, 9]. It is now clear that reasons for and types of anharmonicity in the current–phase dependence differ. The *I–V* characteristics of such junctions are poorly understood. In this work, the effect of the anharmonic current–phase dependence on the *I–V* characteristic is studied with a Josephson junction having an antiferromagnetic spacing. To this end, a term proportional to phase ϕ squared is introduced into the equation of Josephson junction dynamics instead of ϕ, which describes the current through a normal resistance. This is equivalent to introducing a quadratic resistance. The advantage of this model is that the equation for Josephson junction dynamics can be solved analytically. Such a consideration retains generality, since our main goal is to see how the anharmonicity of the current–phase dependence influences *I–V* hysteresis.

BASIC EQUATIONS

It was shown [10] that the nonsinusoidal variation of the supercurrent takes place in the presence of the self-inductance of the junction. In this case, the supercurrent $j(\phi)$ is given by

$$j(\phi) = \sin(\phi - lj(\phi)), \quad (1)$$

where $l = 2\pi LI_c/\Phi_0$ is the dimensionless (normalized) inductance, I_c is the critical Josephson current, and Φ_0 is a fluxon.

With $l \geq 1$, the dependence $j(\phi)$ becomes ambiguous; in the opposite limit, the dependence is unique. Experimental data suggest that this dependence in Josephson junctions with an antiferromagnetic spacing is unique. This indicates that the normalized self-inductance is small, $l \leq 1$. This parameter is evaluated by the formula [10]

$$l = \max_\phi(I(\phi) - I(\pi - \phi))/2I_c. \quad (2)$$

From experimental data [1], we find that the self-inductance normalized varies between 0 and 3: $0 < l < 3$. Note that it increases with decreasing δ. Unfortunately, this relationship may be judged only on a qualitative basis, since a microscopic theory of Josephson junction with antiferromagnetic spacing is lacking. The thickness d of the antiferromagnetic spacing (and, hence, the parameter δ) is involved in the normalized self-inductance l. An analytical relationship between l and d remains to be found. Since l is small, $j(\phi)$ in (1) can be approximated as

$$j(\phi) = \sin\phi - l\sin(2\phi)/2; \quad (3)$$

that is, the presence of a small self-inductance l causes the second harmonic to appear. Later on, instead of Eq. (1) for Josephson junction dynamics, we will use expression (3). First, let us consider a small-capacitance junction. In this limit, the Josephson junction

1063-7842/03/4811-1496$24.00

dynamics is described by the first-order differential equation

$$\dot{\phi} + j(\phi) = i. \tag{4}$$

Here, time is measured in $\Phi_0/2\pi V_{ch}$, where V_{ch} is the characteristic voltage across the junction.

A solution to Eq. (4) has the form

$$\phi = 2\arctan\{[[i^2 - (1 + i^2l^2)]^{1/2}/i(l-1)]\} \times \tan\{(1 + i^2l^2)[i^2 - (1 + i^2l^2)]^{1/2}\tau/2\} + (i(1-l))^{-1}\}. \tag{5}$$

Note that the effect of anharmonicity in this case is quadratically small. The *I–V* characteristic calculated analytically is given by

$$i = (1 + v^2/2)(1 - 3l^2)^{-1}, \tag{6}$$

where $v = V/V_{ch}$ is the voltage normalized to the characteristic voltage across the junction.

Thus, the effect of anharmonicity on the *I–V* characteristic of hysteresis-free Josephson junctions may be neglected.

Now we turn to a large-capacitance junction. In this case, the McCumber parameter is greater than unity [11] and it is therefore of interest to consider a junction with hysteresis. The associated dynamics equation includes a term related to displacement current. An analytical solution can, however, be obtained only with a quadratic resistance [12]. Since the issue is treated on a qualitative basis, we consider an equation in the form

$$\beta\ddot{\phi} + \dot{\phi}^2 + j(\phi) = i. \tag{7}$$

The substitution $\dot{\phi}^2 = z$ transforms Eq. (7) into

$$dz/d\phi + 2az + 2b\sin\phi + 2c\sin 2\phi = 2i/\beta, \tag{8}$$

where $a = b = \beta^{-1}$ and $c = -l/2\beta$.

The stationary solution to this equation is

$$\dot{\phi} = (i + Ae^{-2a\phi} + B\cos(\phi + \arctan(2\beta^{-1})) + C\cos(2\phi + \arctan\beta^{-1}))^{1/2}, \tag{9}$$

where the coefficient A depends on the initial conditions and the term $A\exp(-2a\phi)$ decays.

Thus, the steady-state solution has the form of expression (9) without the term $A\exp(-2a\phi)$. The other coefficients are $B = 2\beta^{-1}/(1 + 2\beta^{-2})^{0.5}$ and $C = c/(1 + \beta^{-2})^{0.5}$. For a large capacitance, Eq. (9) may be solved in quadratures [13]:

$$\phi = 2\arctan(b_0 sc(a_0E^{0.5}(t - t_0)/2.1/\kappa)), \tag{10}$$

where $scu = snu/cnu$ is the Jacobian elliptic function with the argument u, modulus κ, and period T and $E = i - B + C$. Also,

$$a_0^2 = (i - 3C + (B^2 - 8C(i - C))^{0.5})/E,$$
$$b_0^2 = (i - 3C - (B^2 - 8C(i - C))^{0.5})/E. \tag{11}$$

The phase oscillation period is expressed as [14]

$$T = 4K(1/\kappa)/a_0E^{0.5}, \tag{12}$$

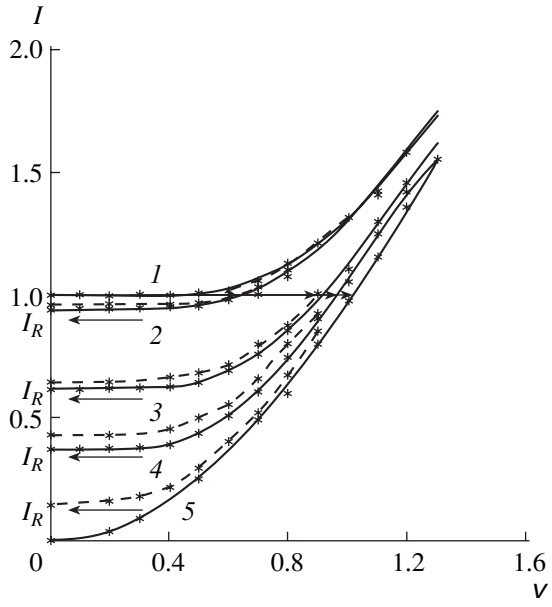
where $\kappa = a_0^2/(a_0^2 - b_0^2) > 1$ is the Jacobian function modulus and $K(1/\kappa)$ is the complete elliptic integral of first kind.

The resistive branch of the *I–V* characteristic is found by the time averaging of the derivative

$$v = \langle\dot{\phi}\rangle = 2\pi/T. \tag{13}$$

The final form of the *I–V* characteristic appears as

$$v = \pi a_0(i - B + C)^{0.5}(2K(1/\kappa))^{-1}. \tag{14}$$



I–V characteristic of a Josephson junction with the harmonic ($l = 0$, solid curves) and anharmonic ($l = 0.2$, dashed curves) phase dependence of the supercurrent for a McCumber parameter β = (*1*) 0, (*2*) 1, (*3*) 20, (*4*) 50, and (*5*) 100.

DISCUSSION

The figure shows the *I–V* characteristics of a Josephson junction for different McCumber parameters β and the inductances $l = 0$ and 0.2. The solid curves correspond to the harmonic phase dependence of the supercurrent ($l = 0$). As the current through the junction grows, the junction switches from the superconducting to resistive state at the critical current I_c irrespective of the McCumber (inductance) parameter β and inductance l (the horizontal arrow pointed rightward from the

point $I = I_c$). With an increase in β, hysteresis in the I–V curve becomes more pronounced; that is, the reverse switching of the junction from the resistive to superconducting state (the horizontal arrows pointed leftward) takes place at a return current I_r, which is lower than the critical current. The return current I_r is described well in the framework of the resistive model [11] by the formula

$$I_r = I_c \frac{4}{\pi \beta^{1/2}}. \quad (15)$$

Formula (15) determines the amount of hysteresis in the I–V characteristic and fits well experimental data. Our calculations in terms of the model with quadratic resistance are in qualitative agreement with this formula: as the McCumber capacitance parameter β decreases from one hundred to zero, the return current grows from zero to the critical current.

The dashed lines correspond to the Josephson junction with the anharmonic dependence ($l = 0.2$). The inclusion of the term $\sin 2\phi$ into the expression for the supercurrent diminishes hysteresis. A low self-inductance of a Josephson junction suppresses the capacitance-related inertial effect and ultimately decreases the amount of hysteresis in the I–V curve. The amount of hysteresis depends on the self-inductance value. In the case of a Josephson junction with an antiferromagnetic spacing, the explicit dependence of the self-inductance on the spacing thickness is unknown because a comprehensive microscopic theory of this object is lacking.

An attempt to work out a microscopic theory has been made in [15] for SNS and SIS structure. When theoretically simulating such structures, one usually considers a potential barrier at the interface between the superconductor and another material. The physical properties of such superlattices depend on the microstructure of interfaces between alternating layers, i.e., on the properties of contacting media (S–S, S–N, S–I, S–F, and S–Sc interfaces, where F and Sc are ferromagnet and semiconductor, respectively) [16, 17]. An explicit expression for the second harmonic of the supercurrent in S–I superlattices was given in [5]. According to [5], the supercurrent in such structures can be generally represented as the sum of all harmonics, $\sin n\phi$, and the amplitude of the second harmonic ($\sin 2\phi$) varies as the interface barrier penetrability squared. Finding the detailed phase dependence of the supercurrent in superconducting structure goes beyond the scope of this work.

Thus, the anharmonic phase dependence of the supercurrent, which is observed in Josephson junctions with an antiferromagnetic spacing and in similar structures, should be taken into account upon analyzing their dynamic properties.

REFERENCES

1. E. Demler, A. J. Berlinsky, C. Kallin, *et al.*, Phys. Rev. Lett. **80**, 2917 (1998).
2. Den Hertog, A. J. Berlinsky, and C. Kallin, Phys. Rev. B **59**, R11645 (1999).
3. K. K. Likharev, Usp. Fiz. Nauk **127**, 185 (1979).
4. I. O. Kulik and A. N. Omelyanchuk, Pis'ma Zh. Tekh. Fiz. **2**, 213 (1975).
5. G. A. Gogadze and A. N. Omelyanchuk, Fiz. Nizk. Temp. **22**, 648 (1996).
6. E. Ilichev, V. Zakosarenko, R. J. Ijsselsteijn, *et al.*, Phys. Rev. B **59**, R11502 (1999).
7. M. Sigrist and T. M. Rice, J. Phys. Soc. Jpn. **61**, 4268 (1992).
8. Yu. Barash and A. Svidzinskiĭ, Zh. Éksp. Teor. Fiz. **111**, 1120 (1997) [JETP **84**, 619 (1997)].
9. Y. Tanaka and S. Kashiwaya, Phys. Rev. B **53**, R11957 (1996).
10. A. A. Zubkov, M. Yu. Kupriyanov, and V. K. Semenov, Fiz. Nizk. Temp. **7**, 1365 (1981) [Sov. J. Low Temp. Phys. **7**, 661 (1981)].
11. K. K. Likharev, *Introduction to the Dynamics of Josephson Junctions* (Nauka, Moscow, 1985).
12. A. Barone and G. Paterno, *Physics and Applications of the Josephson Effect* (Wiley, New York, 1982; Mir, Moscow, 1984).
13. E. Kamke, *Gewöhnliche Differentialgleichungen* (Verlag, Leipzig, 1959; Nauka, Moscow, 1971).
14. *Handbook of Mathematical Functions*, Ed. by M. Abramowitz and I. A. Stegun (National Bureau of Standards, Washington, 1964; Nauka, Moscow, 1979).
15. A. V. Svidzinskiĭ, *Spatially Inhomogeneous Problems in the Theory of Superconductivity* (Nauka, Moscow, 1982).
16. V. M. Gvozdikov, Fiz. Nizk. Temp. **12**, 121 (1986) [Sov. J. Low Temp. Phys. **12**, 69 (1986)].
17. I. N. Askerzade, Zh. Tekh. Fiz. **71**, 141 (2001) [Tech. Phys. **46**, 270 (2001)].

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 11, 2003, p. 1499.

ERRATA

Erratum: "Transition Radiation from an Extended Bunch of Charged Particles" [*Tech. Phys.* 47, 1 (2002)]

B. M. Bolotovskiĭ and A. V. Serov

The plots of the angular dependence of the transition radiation intensity (Figs. 3a–3d) are in error. For the wavelength-to-bunch radius ratio taken in this paper, the intensity maxima at large angles turn out to be less pronounced.

Below, the correct versions of Figs. 3a–3d are given. Note that the discrepancy between the angular distributions of the intensity from the bunch and point source will be much greater for other wavelength-to-bunch radius ratios.

We thank A.P. Potylitsyn who noticed the error.

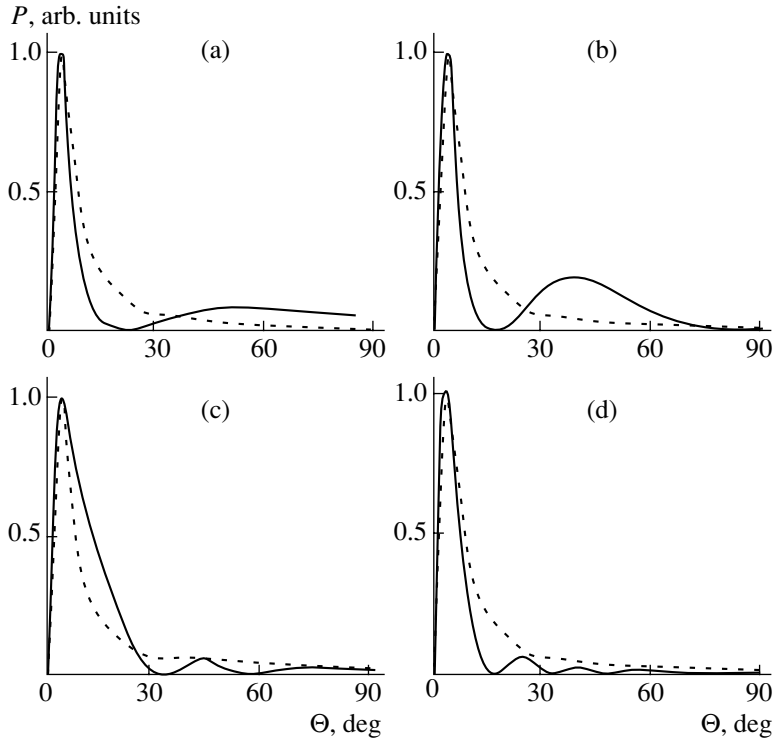


Fig. 3. Angular dependences of the transition radiation intensity at λ/r_0 = (a) 1.5, (b) 0.85, (c) 0.35, and (d) 0.2.

1063-7842/03/4811-1499$24.00